DISCIPLINOLOGY:
REPRESENTATIONS OF THE STRUCTURE OF SCIENCE AND THE HUMANITIES

by

Tyler David Price Brunet

Submitted in partial fulfillment of the
requirements for the degree of
Master of Science

at

Dalhousie University
Halifax, Nova Scotia
April 2016

To Kathryn F. Price & Vincent P. Brunet,

for supporting a young naturalist in all his conflicting interests.

Table of Contents

List of Tables

List of Figures

Abstract

Phylogenetic software and techniques from natural language processing can be applied to the analysis of the structure of academic disciplines. This is possible through a synchronic analysis based on comparisons of the conceptual apparatuses of disciplines, as they are represented in the terminological characteristics of representative discourses. This empirical approach enjoys a superior justificatory status to merely intuitive representations. Finally, this work helps place one of the oldest structuralist problems in the philosophy of science in the scientific context it deserves: "How should we represent the relationship between disciplines?"

List of Abbreviations and Symbols Used

FEG —	Functional Genomics, Evolutionary Biology, Genomics Dataset
FEGC —	Functional Genomics, Evolutionary Biology, Genomics and Chemistry Dataset
HOS —	Hierarchy of Science
NJ —	Neighbour-Joining algorithm
UPGMA —	Unweighted Pair Group Method with Arithmetic Mean
EB —	Encyclopaedia Britanica
SEP —	Stanford Encyclopedia of Philosophy
InPhO —	Indiana Philosophy Ontology
RSPR —	Relativized Sub-tree Prune and Re-graft
SDR —	Subtype Diversity Ratio
TMD —	Topological and Metric Disagreement
CGEB —	Comparative Genomics and Evolutionary Bioinformatics
DPD —	Dalhousie Philosophy Department
inffr —	Informative Fraction
$g(v_i)$ —	Normalized between centrality of v_i
G_x —	Graph object
R^2 —	Regression coefficient
C_i —	Concept set for discipline i
U_{ij} —	Conceptual difference between disciplines i and j
H_{ij} —	Harmonic mean of conceptual divergence

Acknowledgements

Dr. C. Blouin and Dr. W. Ford Doolittle, for supervision, support and encouraging me to pursue this line of inquiry. Dr. Abramson and Dr. Beiko for participation in the defence committee. My parents, K. F. Price and V. P. Brunet for the support that lead me to the point of beginning this chapter of my life. My friends, Dr. Mariscal, Dr. Booth, MacGregor Malloy, Rachelle Gammon, Alaska McMillan, Andy Sears, Devon Scott, Jeremy MacDonald and Brendan Rutherford, for keeping me (mostly) sane through graduate school. Special thanks to Dr. Hleap for technical assistance. Finally the members of CGEB and the Dalhousie Philosophy colloquium series for helpful comments on early work.

Chapter 1

Introduction

To understand complex conceptual phenomenon we often represent our understanding of it in a geometric structure, expecting some things to hold by analogy. The simplest case is a binary opposition, and beyond dichotomy we find a hierarchy of structural representations. To construct these we often rely on our intuitions, and in doing this I believe we often falsify the very phenomena we are attempting to understand. We make the phenomena into a picture of our preexisting intuitions. We should not do this. Instead, as I will show, we should employ the empirical, logical, mathematical, and computational resources available to us when constructing representations of conceptual phenomena. My contribution to this task is at one of the highest levels of generality: the organization of disciplinary knowledge into structural representations of science.

Structural representations of science are ubiquitous. We use them in our everyday speech, often talking about branches of science or neighbouring disciplines (§ 2.5 Common Language Roots). We use them in pedagogy, explaining the complexities of the life sciences by means of a hierarchy of entities—from the minuscule biological molecules up to the bustling societies of which we are but parts (§ 1.1 Linear Hierarchy). And occasionally we use them to construct our identities as researchers, locating ourselves in the hierarchy or on one of the many leaves of the tree of knowledge (§ 1.2 Tree-like Hierarchies).

Our ontology of structural representations of science is quite limited (§ 1.1 Representations of the Structure of Science). Our claims about these structures, our place in them (§ 2.5), are driven mostly by our untutored intuitions about scientific relatedness. Empirical evidence justifying the use of these structures, or the structural representations themselves, is hardly brought to bear on the topic (§ 2.3 Possibility of a Computational Analysis). So our folk- philosophy of disciplinary organization is in need of justification using more than folk-intuitions.

Structural representations of knowledge are canonically construed as either, i) models of our understanding or, ii) ontological depictions of the relationships between disciplines and their worldly domains of study (§ 3.2 The Meaning of Structural Representations of Knowledge). As well as being a source of equivocation in discussions about the structure of knowledge, these two positions do not generally lend themselves to an empirical approach. In order to empirically answer questions about the structures of knowledge I have taken a middle ground between these purely conceptual and ontological stances; I investigate structure using discourse as a proxy (§ 2.1 Justification of a Discursive Analysis).

I explicate the most common structural representations of science: linear hierarchies, branching trees, reticulated trees, networks and rhizomes (§ 1.1-1.5 Representations of the Structure of Science). These five types of representation cover the range of possibilities for structural representations while neglecting trivial ones (Ex/ disjoint and unrelated points, or symmetric polygons). For each structure I present some history of its use, some philosophical analysis of its role in representation, and finally some mathematical and computational methods to generate them. This latter part is done with computational text-analysis tools drawn from natural language processing and structural representation tools from molecular phylogenetics. Combined, these allow the analysis and representation of text data drawn from encyclopaedias and academic journal articles (§ 5. Methods).

After generating representations there remains the tasks of comparing them: to each other statistically, and to our intuitions. Data extracted from academic discourse—in the form of hyperlinks, keywords or whole text—can be forced into any of the common structures (§ 3.2) but nonetheless will dictate which of the range of structures is best. The common structures of knowledge themselves form a kind of hierarchy: each structure is able to account for variation in the data that the former could not. Put another way, each structure is able to represent features of the architecture of knowledge inaccessible to the former (Ex/ there are obvious features of tree-like representations that cannot be captured in any linear ordering).

Through statistical assessment of the fit between different representations, and different types of representations, I show that the linear hierarchical model is incorrigibly naïve, the tree-like models are improvements, and networks are usually best

(§ 5.9 Statistical Evaluation of Structures, § 6 Results). While this trend does not hold in every dataset considered, it does obtain in large or realistically complex ones. There are reasons to prefer a representation other than its ability to precisely fit complex data. Aesthetic and pragmatic costs of complex diagrams often obscure the nature of the data one tried so hard to accurately represent. So this computational analysis will throughout be paired with a philosophical one (§ 2.4).

Placing this work in relation to other disciplines of science is itself a philosophical difficulty (§ 7, Two Analogies: 1916 and 1977). Were the analysis here directed only at biological literature one might well call it biological informatics. While the title informatics fits, the application of this approach to discourse clearly enjoys bibliometrics status. Beyond methodology, my concern is to explicate the debate about the representation of structure of science from an empirical standpoint; clearly a pursuit in the philosophy of science. Perhaps it is best to follow the title of Paul Thagard's 1988 book (Thagard 1993), in writing if not in spirit, and say that this work is some sort of "Computational Philosophy of Science".

Representations of the Structure of Science

In the following portion of this chapter I elaborate on the range of possible structural representations of science, noting their historical underpinnings, fundamental characteristics and role in representations of different types.

1.1 Linear Hierarchy

Definition: an arrangement of entities along a spectrum or axis. May be polarized to indicate the extent to which entities possess a property.

The most commonly held belief about academic organization is as follows: concepts are grouped into disciplines, and disciplines are arranged in a hierarchy[1]. Take, for example, a hierarchy beginning at mathematics or logic and progressing through

[1] It has been noted by Dr. W. F. Doolittle that this structural arrangement might be more naturally termed a spectrum than a hierarchy. Indeed, a spectrum does not have the same connotation of privileging one pole over the other. Nonetheless many lay conceptions of the hierarchy of science do come with privilege given to physics, or mathematics, for example.

physics and psychology to sociological and anthropological[2] study. The linear arrangement of disciplines is often conceived as reflecting an increasing complexity of subject matter or decreasing consensus amongst researchers—an idea that is usually attributed to Auguste Comte (1835).This intuitive model of the organization of knowledge has received a great deal of support and criticisms—embodied in the currently unfashionable "hard-soft" distinction amongst the sciences.

Daniele Fanelli and Wolfgang Glänzel (2013) even recently claimed to have "conclusively proved" that a linear-hierarchy—what they call a Hierarchy of Science or HOS— is the structure of scientific organization. They attribute this trend—observed in a series of metrics of consensus applied to articles grouped by discipline—to a gradual change in complexity of subject matter up the hierarchy. This general Comtian presumption seems problematic in three ways. Firstly, while few would disagree that arrangements of matter increase in complexity (however operationalized) with the scale at which they are observed, it is highly questionable whether such an increase can be directly related to fields of study—many of which incorporate data from a range of scales and degrees of physical complexity (Dupré 1983). Secondly, it seems at least possible, if not likely, that a discipline should be concerned with highly complex entities, yet nonetheless researchers in the field tend to reach consensus and vice versa; perhaps because they happen to use similar methods. Finally, it is hard to ignore the danger of reductionism that comes along with any linear ranking: aligning theoretical or explanatory scales of complexity with physical ones (see Kitcher 1984).

Regardless of our tastes for reductionism, we can question whether or not something as abstract as a structural representation of knowledge can even be "conclusively proved". Indeed, all of the structural representations I will discuss here come with their own set of techniques for obtaining them from data, techniques that are more or less open ended as regards the kind of structures that could be obtained therefrom. The metrics applied by Fanelli and Glanzel (2013) have two possibilities: linear hierarchy or not (see methods § 5.8 for discussion of a relatively open ended approach).

[2] This point is commonly unrecognized anthropocentrism.

1.2 Tree-like Hierarchy

Definition: an arrangement of entities along a set of paths of successive bifurcations. The paths may be undirected, giving an unrooted tree, or all directed away from a given entity, giving a rooted tree. Depth within the tree may indicate inclusion within deeper entities, giving an inclusive hierarchy, or difference from deeper entities, giving an exclusive hierarchy.

Examining the bureaucratic organization of academic institutions we commonly find a tree-like (AKA arborescent) inclusive hierarchy; e.g. philosophy and social science are nested under humanities with biochemistry and histology under natural sciences. As well as for its likeness to institutional organization, one might prefer such a structure on historical grounds: the branching of disciplines off of ancestors. Perhaps from a western perspective we might imagine placing the apex—root—of the tree somewhere in the geometry or philosophy of ancient Greece. Circa 1750[3], Jean le Rond d'Alembert and Denis Diderot[4] published a tree-like classification of human knowledge ("Système Figuré des Connoissances Humane") in the 'Encyclopédie, ou dictionnaire raisonné des sciences, des arts et des métiers' (Dederot and d'Alembert 1772). In this tree of knowledge the branching pattern was explicitly organized by human reason as opposed to by nature or theological dicta (Darnton 2009). Interestingly, the tree-like representation is almost a perfect binary branching tree, aside from one reticulation in a branch that appears to be connected to both "Narrative" and "Drama".

Tree-like classifications are also often presumed to align with levels of emergence or scales of complexity. In his attack on reductionism—and thus on the linear-hierarchy—Dupré notes the slight superiority of a treelike-hierarchy.

> One striking oversimplification is that macro-physics has been entirely excluded. Molecules do not only go to make up cells, but also many complex non-living structures. At least science should be viewed as a branched tree to accommodate gravitation, electromagnetism, cosmology, geology, and so on; though it is in fact not clear how even these branches should be rooted in microphysics. (Dupré 1983)

[3] 50 years before Comte's hierarchy, between 1751 and 1772.

[4] For analysis of Diderot's philosophical contributions see Brewer (2006).

It may be a ubiquitous human tendency[5] to classify both natural and conceptual things in groups subordinate to groups (as opposed to disjoint classifications). Nonetheless, unlike in biological systematics where evolution gives us a natural reason to expect broadly tree-like categorizations, disciplinary organization has no such unifying explanation. The relations of ancestry and descent between organisms, genomes, instantiate a process that tends to generate variants that are differ from their parents by a small margin over long periods of time. Thus one can, in many cases, treat differences between genomes as approximating times or orders of divergence amongst biological entities. No such assumptions about change are *as* universally acceptable when the entities in question are disciplines, texts, or other knowledge resources. If horizontal gene transfer is a problem for grand-scale trees in genomics, then *horizontal concept transfer*[6] is an even more serious problem for even less grand-scale trees of conceptual apparatuses. Indeed, while linguistic applications of phylogenetics, such as glottochronology, may have some merit (see Heggarty 2006), the attempt to assign historical meaning to tree-like analyses of disciplines (within a language) is a different matter.

I believe this anti-genealogical point will hold regardless of whether we treat disciplines as historical entities, conceptual structures, or organizations of the emergence of nature. Likewise, we have no more reason to suppose that a treelike-hierarchy should correspond to the emergence or complexity of nature than a linear-hierarchy. The best we can expect in any case is that one will better suit our needs and be more or less reflective of the organization present in our discourse. In general, in lieu of some privileged access to metaphysical or ontological reasons for disciplinary organization, the patterns we observe in structural representations of science should instead depend on our writing practices and the organization we *impose* on our discourse.

1.3 Reticulated Tree-like Hierarchy

Definition: an arrangement of entities along paths with bifurcations and intersections. All paths are directed away from a given entity, giving a reticulated

[5] Indeed, in biological systematics it was a *discovery* that organisms could be classified this way.

[6] While the horizontal transfer of concepts between languages is properly called 'word borrowing' (Heggarty 2006), such transfers between other knowledge resources, disciplines, etc., seem deserving of a coinage.

rooted tree. The extent of intersection between entities may indicate the extent to which entities share a property.

Consider again Dupré's (1983) speculation on the structural representations of science. When he offers up the problematic place of Ecology for explanatory reductionism he also offers up a challenge to consider structural representations beyond the branching tree,

> Ecology is in fact a particularly interesting case... Not only do ecological systems typically combine elements from at least three levels: multicellular organisms, single cells, and molecules (as nutrients in the environment), but the understanding of such systems also involves such factors as climate and geology, which would have to be assigned to parallel branches [of the tree of knowledge]. Whether a plausible model could be constructed to take account of all these complexities is a question I shall leave open, though it would certainly not be as simple a task as is sometimes supposed. (Dupré 1983)

Ecology is not alone. Many factors serve to complicate the hierarchical picture, including some post-modern analyses of the genesis of disciplines highlighting the sharing of concepts and methods during their growth (Lenoir 1993; Foucault 1970; Keller 1991). This type of analysis is often referred to as genealogical, by analogy, and after Foucault's use and expansion of a similar concept in Nietzsche's 1887 Genealogy of Morals (Nietzsche 1956). These analyses bring to light the frequency of interdisciplinary mechanisms of disciplinary change and transfer. Instances of horizontal transfer of concepts and methods between disciplines suggest at least some reticulation in any hierarchical structure of disciplines, and cast further doubt on the rigid separation of academic departments—the conceptual apparatuses of those departments—by emphasizing the essentially interdisciplinary (and non-disciplinary) nature of their origin.

Complication of the search for a structure of disciplinary relations does not stop at genealogy. Many theorists (see Hoskin & Macve 1986) up to the present have examined disciplines as political structures—equipped with a power-dynamic and various modes through which knowledge, and thus discourse, is generated and authorized.

This politico-economic view suggests that concept-association in and between disciplines is often independent of theoretical difference, and depends instead on a series of historic contingencies and human power struggles for intellectual authority and financial support.

The notion of disciplines combining information from “parallel branches”, “theoretically independent” concept association, along with the “horizontal transfer” of concepts, are surely tree-violating aspects of the representation of the structure of disciplines: they induce reticulation. Indeed, devising a plausible model of science that accounts for these events is both difficult to do (as Dupré pointed out in the case of Ecology), and difficult to motivate. A parallel problem has emerged in molecular phylogenetics with the growing knowledge of the extent of horizontal transfer of genes between organisms: likewise violating the strict gene trees representing standard Darwinian descent with modification. I believe the analogy here indicates what such a plausible model might look like, *since the software developed to model reticulation in gene trees can be transferred to reticulations of a more conceptual nature* (see § 3.1 General Notion of an Empirical Structure of Knowledge; § 5.10 Visualization)

One could argue that I have taken too many liberties in naming structures as tree-like instead of, i) wholly a network with a root or, ii) wholly a tree with insignificant reticulations. That issue is a current matter of debate in the philosophy of phylogenetics surrounding the tree of life. Consider the position of David A. Morrison (2014) on the issue in that domain,

> [A] “tree with reticulations is a network...therefore a network will be a better metaphor, model, and heuristic for phylogenetics, in the sense that it will be more inclusive and more powerful. This distinction between tree and network in the face of reticulations is not a semantic one. The tree metaphor/model/heuristic pre-supposes tree-like data, whereas the network allows the data to determine the tree-likeness of the metaphor/model/heuristic—some networks are more tree-like than are others. (Morrison, 2014)

While not entirely semantic, the distinction seems to be one of preference when it comes to terminology, and prudence when it comes to analysis. Non-treelike data can be forced into representations as trees, and *vice versa*. So long as the methodology involved in generating a structure *allows the data to determine the tree-likeness of the representation*, it seems arbitrary from a modelling perspective whether the representation is described as a ‘tree with reticulations’ or a ‘tree-like network’ (presuming it has reticulations at all). Now, it may be true that the ‘tree with reticulations’ view has a tendency to undervalue those reticulations when treated as a metaphor or heuristic, but it is hardly the fault of the representation or model that its proper name is metonymic with one having less rhetorical force.

1.4 Networks and Graphs

Definition of Network: an arrangement of entities along paths with bifurcations and intersections. Not all paths are directed away from any given node. Extent of intersection between entities may indicate the extent to which entities share a property.

Definition of Graph: a mathematical object consisting of two sets called its nodes and its edges. Nodes may be connected by edges given some rule, and both nodes and edges may possess additional attributes identifying them as particular entities.

If we entirely reject the assumption that our representation of the structure of disciplines need be hierarchical we arrive at the notion of networks of disciplines, i.e. a reticulated non-hierarchical tree. The mathematical or computational corollary of a network of relations between individuals is a graph with nodes representing objects and edges representing relations. Connections between nodes are made based on some set of rules, and attributes can be added such as edge weights or node labels[7]. For example, we might have a rule that we connect two nodes representing people in a social group just in case they are friends on some form of social media, thus generating a social network or friendship graph. We might also have a rule that we connect two nodes representing publications in a database if they share a common author, keyword or if one is cited by the other—thus generating an authorship, conceptual or citation graph of the database. It is of course possible to have more than one

[7] Indeed, all preceding structures *can* be defined as particular types of graphs, but it would be unfair in the context of the origin and use of these structure to reduce them all to the mathematically precise definitions given in graph theory.

rule involved in generating a graph, more than one type of node (publications and authors) and more than one type of edge (citation and co-authorship). (For some indication of the problems to which concept-graphs can be applied see § 2.3 The Possibility of a Computational Analysis).

Unfortunately for historical context, since networks and graphs are *the most general* type of structural representation involving lines (edges), points (nodes) and labels (attributes), it is difficult to disentangle its history from that of less general structures. John F. Sowa (2006), a theorist engaged with the use of semantic networks within artificial intelligence, for instance pinpoints the "oldest known semantic network" in precisely the same place I pinpoint the oldest tree of knowledge: the *Tree of Porphyry*.

Indeed, while historical context is perhaps lacking, logical analysis is surely not. Plenty of logical structures count as network type representations of conceptual information, such as Frege's *Begriffsschrift* (concept-writing) or Pierce's relational graphs (Sowa 2006, Frege 1879). But in the present connection, I will confine myself to large scale quantifiable relations between discourse-level conceptual objects, and not to formalisms of more sentence-level aspects of conceptual relatedness. So I will confine my comments on the diversity of network representations to those structures with more general applicability.

One such structure worthy of historical note is the citation graphs of EigenFactor®, used for "mapping the structure of academic research" (EigenFactor® 2015). Perhaps more widely known for their journal rating system, EigenFactor® has also contributed to the visualization of citation data (Accessible from EigenFactor® citation graph). Network representations are powerful tools to be sure, but only ever as powerful as the data used to generate them and the rules that establish their connections. Citations are indeed a powerful form of data, but they are but one of many different kinds of data that one could use to draw connections, and so I do not think they deserve special status as the decisive factor in determining *the* structure of knowledge.

Networks and graphs have the following virtues. There is a convergence of ontological, discursive and conceptual representational goals. Networks are preferred as ontological representation for the same reason they are preferred as conceptual or discursive representations: they are capable of representing complex interdependencies. If one is not inclined towards viewing the natural world, texts or disciplines as having a simple, independent existence, then networks meet the desiderata.

Unlike trees or linear representations, networks are necessarily non-hierarchical. Indeed, the only property of networks that resembles hierarchy is "directedness", or the direction of edges within a graph. While the combination of many directed edges within a graph can impact the "flow", the combination of these individual directions need not induce any global network polarization (see EigenFactor map equation method for more precise notion of probabilistic flow through a graph). This lack of global polarization also makes reading unjustified historical narratives into networks at least unlikely if not nonsensical; not to say that historical narratives are not involved in explanations of network characteristics.

As well as lacking certain philosophically suspect features of less general representations, networks are also amenable to additional analysis on the basis of their modularity. While there are a multitude of distinct methods of computing the modularity of groups within a network, an informal motivation and definition will suffice for the present purpose.

> [I]f the number of edges between groups is significantly less than we expect by chance, or equivalent if the number within groups is significantly more, then it is reasonable to conclude that something interesting is going on.
>
> ...
>
> The modularity is, up to a multiplicative constant, the number of edges falling within groups minus the expected number in an equivalent network with edges placed at random. (Newman 2006)

So while of course the interest one might have in a module (AKA 'group' or 'community') will depend on our confidence in the procedure used to generate the graph, one can nonetheless use the intuitiveness of the modules defined on a graph as an indication of the confidence one ought to have in a graph structure. Moreover, since in the present connection the *nodes in the graph* are the same as the *leaves on the tree*, one can compare the modularity within a graph to the domain architecture of the tree (see § 5.7, 6). So as well as being a structure in its own right networks—graphs—can be used as tools for the analysis of other structures.

1.5 Rhizomes

Definition: a graph or sub-graph in which every node is connected to every other node by at least one edge (directly). There may be more than one edge between nodes and edges may have different attributes, resulting in local distinctions between node connectivity. Known as a maximal-clique when the largest such sub-graph in a graph.

Trees and webs are not the only biological metaphors proposed as models of linguistic and conceptual structure. Deleuze and Guattari (1980) introduced the concept of a rhizome to explain an anti-genealogical, non-treelike, structure organizing a variety of things from linguistic phenomena to symbiotic relationships between animals. In *A Thousand Plateaus* Deleuze and Guattari outline some approximate characteristics of rhizomes. Of interest here in distinguishing rhizomes from networks is the principle of connection.

> [A]ny point of a rhizome can be connected to anything other, and must be. This is very different from the tree or root, which plots a point, fixes an order. Deleuze & Guattari (1980)

The principle of connection clearly establishes rhizomes as a limiting case of networks and graphs—the case where the entire network or graph is a *maximal clique*. A clique is a complete sub-graph, i.e. a subgraph where every node is connected to every other, and a maximal clique is just a clique where it is impossible to add any other nodes from the graph into the clique. Identifying such structures has been problematic in graph theory—due to the computational burden of doing so—but is also fruitful within applications of graph theory. Depending on how they are constructed, rhizomes can have some interesting properties. Bapteste et al (2012), for instance, created networks from BLAST[8] searches, drawing edges only between high scoring homologues, and used the presence of maximal cliques as a proxy for those sub-graphs that were amenable to more traditional phylogenetic analysis.

Wikipedia hyperlinks are not rhizomatic in general—there are pages to which no other page is directly hyperlinked (Zlatić et al. 2006). Nonetheless, we expect

[8] BLAST is a phylogenetics application that searches a database for homologues—sequences related by descent—when given a query sequence

a higher degree of connection amongst the large and interconnected pages typically devoted to encyclopedic entries on academic disciplines. No discipline is an island. Moreover, if a metric is employed that accounts for more than hyperlinks between pages, or co-occurrence of word tokens, when drawing connections between points, we *might* expect the entirety of Wikipedia to demonstrate such a structure (see Table 4 for demonstration of the rhizomatic character of a set of canonical disciplines). Masucci et al. (2011) actually take this type of structure for language as a given at the commencement of their study into the modularity and topology of the 'semantic space' of Wikipedia.

The important distinctions between rhizomes and networks are firstly, but not only, historical: rhizomes originated in and had the most uptake amongst continental philosophers, while graph and network are the popular terms amongst mathematicians and scientists respectively. Nonetheless the original use of 'rhizome' as a descriptive term for structural connections is conceptual: Deleuze and Guattari's use of the term is explicitly with reference to the organization of concepts. The more abstract, mathematical, or scientific use of 'fully connected network' or 'maximal clique graph' fail to have this necessarily linguistic, conceptual, connotation.

The Hierarchy

This concludes the discussion of those potential structures for knowledge that I believed were worthy of a full length treatment. Nonetheless, for the sake of completeness I believe I should very briefly outline the entire hierarchy of structural representations. In general, a structural representation must have some dimension and connectivity, so a hierarchy can be established on the basis of increases in these features. Trivially we can imagine a *singularity* view of knowledge were all facts are represented as co-located in a zero-dimensional point without any potential for difference or connectivity. We can then add more such singularities and obtain a *binary*, *trinary*, etc., representation of wholly distinct domains of knowledge—separate and each of dimension zero, these representation are necessarily disorganized. Adding a full dimension brings us to *linear* representations, these being hierarchical and having the potential for connections to adjacent points on the line. When we move to two dimensions and a connectivity of maximum 3 we obtain *trees*. Increasing the connectivity indefinitely, relaxing all constraints on connectivity, gives *reticulated trees* and *networks.* When min connectivity is equal to the number of elements of knowledge we obtain *rhizomes.* Finally, if we allow representations in three dimensions (that necessarily require three, since any tree can be represented in three if one pleases), and drop the requirement for connectivity, we step into the world of *3D-clustering* approaches.

1.6 Expectations and Limitations

Already we have a series of hypotheses available for structural representations: the domains of humanities, social and natural sciences, linear hierarchies, treelike hierarchies, genealogical networks, and rhizomes; a significant improvement over the null-hierarchy dichotomy of Fanelli and Glänzel (2013). Yet we expect a different structure of disciplinary relations depending on whether we examine the historical genesis of a set of disciplines—the genealogy of its practitioners and concepts—or we perform comparisons on current discourses. Genealogy has been corrupted by inter-discipline concept transfers. This interest relativity is especially important to observe in this case. I am not attempting to conduct any sort of "computational genealogy" of disciplines—although that is not to say that various analyses might not reflect aspects of genealogy as they are presented in discourse[9].

The analysis presented below is based on synchronic comparisons of the current conceptual apparatus of disciplines, as they are represented in the terminological characteristics of representative discourses. Anything more (genealogical, historical, ontological, or geographic) requires additional theoretical assumptions.

Setting aside the idealized notion that any conceptual structure will exactly follow a hierarchy or the branching pattern of a tree, we can instead consider the degree to which we expect them to be tree-like. As its name perhaps suggests, tree-likeness is not itself an exact notion, so an exact method has been developed (§ 5.9 Statistical Evaluation of Structures) to assess the distortion of data when a tree structure is imposed upon it.

The genealogy of practitioners might be expected to tend more towards a treelike

[9] An early test analysis was done on the set of Wikipedia pages for countries, this analysis showed domains, clusters, corresponding not only to geographical proximity but also historical relatedness (the presence or absence of colonialism). Nonetheless, this is no grounds for claiming the analysis itself was "historical", "geographic", or "colonial".

structure than would a conceptual map, but reticulation can occur from horizontal practitioner transfer as much as by horizontal concept transfer. Researchers rise up through an education system and gain training from predecessors in a particular field, and work roughly within the bounds of that field, employing its terms and methods. On occasion they branch off from their cohort and begin working within a new or underdeveloped field. *Here the metaphor of 'branching' takes on a quite literal character.* Nonetheless, researchers often move between fields and can occupy a position in many otherwise independent fields (meanwhile transferring concepts horizontally between discourses). Thus even this genealogy of practitioners cannot avoid some blurring and reticulation.

The conceptual apparatus of disciplines should likewise be expected to show significant reticulation. Researchers do not devise an entirely new set of methods, terms, and concepts when they begin working in a new field—when a new discipline begins to be conceptualized—nor do they merely modify the conceptual apparatus of some ancestral discipline. Instead they borrow, steal, or reinvent many components from other neighbouring or distant fields of study[10].

Foucault, in "L'Archéologie du Savoir" (1969), addresses the topic of concomitance or "the configuration of statements from quite different domains to different types of discourse" (Lenoir 1993, pg.74).

> Thus the field of concomitance of the Natural History of the period of Linnaeus and Buffon is defined by a number of relations with cosmology, the history of the earth, philosophy, theology, scripture and biblical exegesis, mathematics (in the very general form of a science of order); and all these relations distinguish it from both the discourse of the sixteenth century naturalists and of the nineteenth century biologists. (Foucault 1969)

One can already see in this kind of analysis the network-like characteristic of such

[10] It is important to emphasize again the metaphors we use to characterise differences in subject matter, since from them it is apparent that our folk-philosophical conception of disciplines is well engrained.

associations, and, in the style[11] of Foucault, a purely discourse-based method of analyzing the concomitant disciplines associated with a given discipline of interest. And this concomitance is just the first source of non-treelike descent with modification. I expect that any discourse-based analysis that is sufficiently rich in its appreciation of the complex interactions between disciplines will need to account for reticulation and "horizontal transfer" when constructing a representation of the structure of disciplines. That is, I expect that more complex sets of discourse will generally require networks—an expectation that has been borne out, at least, within the results presented here (see § 6.).

When it comes to investigating concepts by means of text-tokens of words, it is prudent to acknowledge from the onset that a few assumptions are required to begin such analysis. First is the assumption that concepts can be roughly identified by the occurrence of named entities. The primary complication here is that, homographs, heteronyms and homonyms cannot be identified without context specific information that is not present in the mere token (for a more complete list of factors that complicate this analysis and their effect on measures of conceptual divergence see Table 1). Second is the assumption that co-occurrence of terms can roughly be identified with their correlation. The complication is that often co-occurrence is random or actually indicates dissonance; although there is no denying that dissonant concepts are nonetheless related by their dissonance. Nonetheless, it should be noted that if word tokens are bad proxies for concept use, then it is not clear to me what a *good* proxy for concept use could be.

Neither the local nor global structure of our discourse can or should be determined a priori—instead we should be looking for an approach that is both empirical and open to a wide range of potential structures. Some phylogenetics software (like SplitsTree) falls into this category of open-ended approaches that allow one to obtain many different types of structural representations from the same data.

It was precisely the absence of possible alternative structures that I take to be the greatest mark against Fanelli and Glänzel (2013)—their analyses applied linear metrics to journal articles to see if they would display a certain linear ordering (as

[11] Foucault had a particular interest in confining his philosophical remarks on history to what could be gleaned directly from written texts, and in this sense a text based analysis is, in a way, in keeping with his method.

opposed to a binary ordering or no ordering at all), and thus considered only a small portion of the range of possible structural hypotheses.

Indeed a variety of confounding issues obtain even in the search for an exact terminology for structural representations. In the case of any hierarchical arrangement Ernst Mayr points out that 'hierarchy' is in fact ambiguous between two distinct notions.

> Most classifications, whether of inanimate objects or of organisms, are hierarchical. There are 'higher' and 'lower' categories, there are higher and lower ranks. What is usually overlooked is that the use of the term 'hierarchy' is ambiguous, and that two fundamentally different kinds of arrangements have been designated as hierarchical. A hierarchy can be either exclusive or inclusive. (Mayr 1982, p. 205)

An exclusive hierarchy is one where the lower members are not included as parts of higher ones—such as the ordering of military ranks—whereas an inclusive hierarchy does have its lower members included as parts of its higher ones. Certainly the conceptual apparatuses of disciplines are neither: they are neither entirely nested within one another, nor are they entirely disjoint domains. So if the relationship between disciplines can be considered hierarchical at all, it will have to be a varied mixture of exclusive and inclusive types. But (as will be described in § 5.5) the *very fact that there exists variation in the overlap of the conceptual apparatuses of disciplines provides a means of empirically constructing hierarchical representations.*

Chapter 2

Motivation

In this chapter I will attempt to clarify in turn the roles and interdependencies of non-conceptual facts, discourses, computation, philosophy and common-language within analyses of the representation of the structure of science.

2.1 Material Aspects of Disciplines

Timothy Lenoir, in The *Discipline of Nature and the Nature of Disciplines* (Lenoir 1993), addresses the role of disciplines in stabilizing scientific practice.

> Within this complex of issues generated by the disunity of science, disciplines emerge as a crucial site; for as laboratories and sites of apprenticeship are essential for organizing and reinforcing the economies of skill necessary for conducting science locally, disciplines are the structures in which these skills are assembled, intertwined with other diverse elements, and reproduced as a coherent ensemble suitable for the conduct of stable scientific practice more globally. (Lenoir 1993)

Indeed, I shy away from introducing Lenoir's opposition of the local-laboratory and global-discipline into my own analysis, since I doubt that we can evidence a separation between local-laboratory-discourse and global-discipline-discourse (see § 2.2 Justification for a Discursive Analysis). But at least such characterizations of scientific disciplines—by opposition with the laboratory component of training—enjoin us to take seriously the non-conceptual characteristics of disciplines, such as practitioners and resources.

In general, we can be confident that every extant[1] discipline will have practitioners: working members of a community of academics who strive either for the elaboration of knowledge within the purview of the discipline[2], or who seek a complete reorganization or elimination of the boundaries of knowledge therein. It is neither necessary that each academic practitioner be categorized by their place within a specific discipline, nor that an academic from some given discipline or other be incapable of also participating in a quite distantly related discipline[3].

Every discipline will have a set of resources that help further the work of practitioners therein. Resources are often introduced from outside and specialized for the tasks specific to the practitioners of a given group, or a discipline forms around the use of a resource, instead of a resource being acquired to suit the needs of a discipline[4]. This dependence on resources is most obvious in the case of the physical apparatuses of experimental sciences: everything from Petri dishes to the Large Hadron Collider are reminders of the material and financial resource dependence of scientific disciplines. Yet the dependence of a disciplinary community on grants, access to information, graduate students and finally a working environment[5] are universal.

Together these *material* considerations highlight some of the common points raised by those who criticize a theory-dominated approach to the study of disciplines.

[1] Certainly there are disciplines that have died out, or been so heavily supplanted by their modernizations that they truly possess no practitioners, resources, etc. In these cases we are dealing with an historical discipline, or, a living fragment of an historical discipline.

[2] Researchers often have no prior intention of 'furthering a discipline' with their research, but the process of categorization following the consumption of their work often ensures that it is interpreted as such, i.e. it is often only long after research has taken place that some work becomes canonized as a major contribution to some field or other. Consider the "foundational" work on Turing machines in light of the apparatus of modern computer science.

[3] Considering that disciplines often get their names by having diverged or merged with other disciplines or branches of inquiry, it seems overly simplistic to suppose that any two disciplines should be completely opposed or cut off from each other. Indeed, I believe this is sufficient motivation to consider the relationships between disciplines as greater and lesser degrees of relatedness rather than as opposition or contradiction.

[4] Aquaculture being an example.

[5] Resource is such a general category. We might imagine the more sophisticated social theorists interjecting that academic life as a whole, and thus disciplinary structure, depend on a whole political and economic climate. The interesting question here is how and where the specifics of the politico-economic climate influence changes in which disciplines are fostered in an academic community. Western liberal-democracy was sufficient to allow 'Religious Studies' in public schools and universities to include Taoism, whereas its practice elsewhere is undeniably more parochial.

2.2 Justification for a Discursive Analysis

The point of departure, from a merely verbal analysis, is concept use within a discourse. It seems quite natural to identify disciplines by the concepts used by their practitioners instead of, perhaps, by their tools and other accoutrements. But as we have seen above (§ 1.2-1.5) there are plenty of reasons to see it as necessary to include other entities than concepts in our analysis of disciplines. So we have good reason to believe that a merely conceptual (theory-dominated) analysis of disciplines is insufficient, but I also think we have good reasons to believe that a discourse-based analysis is not merely conceptual.

Rejections of a theory-dominated approach to studying disciplinary relations often cite the overwhelming influence of non-conceptual (i.e. material) factors—practitioners, political climates, and resources (Lenoir 1993; Keller 1991; Hoskin and Macve 1986). Yet a clear line between conceptual and non-conceptual influences does not exist within discourses: names of practitioners, theoretical concepts, descriptions of resources and methods all appear within the discourses associated with disciplines. So analyses based on discourse cannot necessarily be derided as merely theoretical investigations—devoid of practical and historical weight—since non-theoretical terms form a central part of discourse.

In short, it is not as if discourse, consisting only of words, described only concepts and left out mention of the non-conceptual foundations of disciplinary life. Non-conceptual terms appear within and help to structure disciplinary discourse. Thus they, as much as theory, aid in the quantification of difference between disciplines.

2.3 The Possibility of a Computational Analysis

There is plenty of precedent for application of structural metaphors as models of scientific data. The case here examined is no different in that respect—even given the obviously historical, political, philosophical, and conceptual underpinnings of disciplinary organization.

We are all exposed, dogmatically at first, to the basic concepts and particulars of practice that are presumed to characterize disciplines—the divisions of the school systems ensure this. Yet there are examples abound of interesting phenomena that do not fit nicely into the naïve systems[6] thus far considered. The question of 'the subject matter of discipline X', then becomes a matter not of rigid disciplinary boundaries—separating disciplines from each other and so too their subdisciplines—but of distributions of lay concepts, terms of art specific to disciplines, and a network of shared and interrelated concepts.

On this last point—the network of concepts—one can easily interpret such a phrase as mere metaphor. Yet with advancements in phylogenetic tree construction, network analysis, text-mining and natural language processing, we can now begin investigating to what degree such a metaphor serves as an interesting and useful aim of computational analysis. What would an empirical tree of knowledge look like? What networks can be drawn from the interrelation of concept use amongst disciplines? What groups can be defined or clarified by searching through such networks? And what conclusions can be drawn by statistically evaluating these networks? *All are questions now answerable on grounds more empirical than philosophical or philological.*

The mixed empirical-philosophical nature of the questions here posed is certainly reminiscent of the experimental philosophy movement—at least, in as much as philosophy is not usually empirical and computational text-analysis is. Yet a quasi-empirical approach to philosophical questions may be all they have in common: experimental philosophers have traditionally be concerned with uncovering people's moral intuitions, underlying facts about consciousness, or their beliefs about reference (Knobe and Nichols 2008). The analysis here has far different applications.

2.4 The Utility of a Philosophical Analysis

The notion of a science of science—or at least a computer science of the structure of science—ought to invoke in us the same visceral opposition to inbreeding that we experience in familial relations; except in this case it is an inbreeding of intuitions. There is a danger of scientific intuitions about the structure of science becoming self-justifying and serving as the basis of quasi-empirical models of the structure of science. It is precisely this limitation in the number of possible results of a model

[6] An instance of quantum mechanics being required for explanations of a model of photosynthesis in grass plants comes to mind.

of the structure of science (exemplified by Fanelli and Glänzel (2013) and again by Colin Allen in the *Indiana Philosophy Ontology project* in § 4.1) that has lead to my current more open-ended approach.

Another source of opposition (to philosophy of science generally) is embodied in the phrase, "Philosophy of science is as useful to scientists as ornithology is to birds", that is usually attributed to Richard Feynman. I take this analogy to roughly express the following criticism: scientists are doing just fine on their own, so philosophers don't really need to step in and do anything[7]. But this argument from analogy is surely flawed on both ends. Conservation of bird species requires ornithological expertise, and scientists are certainly not self sufficient even now[8]. If we are going to properly conduct interdisciplinary research we ought to do so with more than our untutored intuitions about scientific relatedness.

The problem of the structure of knowledge began in philosophy, was specialized as the structure of science by philosophers of science, and most recently taken up by bibliometricians and scientometricians again as the web or hierarchy of the sciences. That an old problem has seen new variations, had new techniques brought to bear upon it, is certainly no reason to turn away from the context of the original formulation: fostering an understanding of knowledge writ large.

2.5 Common-Language Roots for Structures of Knowledge

Were it not for the structural implications of our everyday parlance one would be tempted to think of structural representations of knowledge as nothing more than a simplification of our understanding, a cartoon sketch of our understanding, a bibliometricist's fancy, or best fit for creating clever info-graphics to place on the front of secondary school Science textbooks.

[7] Of course, we could also interpret this quote as attempting to solidify the same respect and grandeur for ornithology as is possessed by the preceding two thousand years of philosophy.

[8] For example, consider the recent case of ENCODE (Doolittle et al. 2014), a multimillion dollar research enterprise devoted to finding all the functional elements in the human genome. ENCODE suffered serious backlash, from scientists and philosophers, about the ignorance of the project leaders of the philosophical analysis of the concept of function. I see this as a case where philosophers have successfully critiqued a major scientific project and showed that the working scientists were not self sufficient: they should have outsourced their conceptual analysis of the notion of function they were implicitly working with.

But we do employ structure-of-knowledge talk daily, and to various degrees of sophistication. Certainly nobody moves though the education system without hearing the likes of,

> What *branch* of science does she work in?
>
> Then she progressed to work in a *neighbouring* discipline.
>
> No, entomology has nothing to do with words, it is a *sub*discipline of biology!
>
> Well...I guess my thesis is kind of *inter*disciplinary.

One will also hear the terms ‘multi-disciplinary’, ‘transdisciplinary’ and even ‘nondisciplinary’ being tossed about.

Each of these references to disciplinary organization comes equipped with presuppositions about structure, or at least has implicit, characteristic, structural implications. Branching can obviously be likened to a tree model (hierarchical or otherwise), neighbouring and interdisciplinarity can be taken at least to imply notions of relative differences in distance (another structural metaphor) amongst disciplines, and notions of super- and sub-disciplines have obvious hierarchical connotations. Given these roots, I believe it would be a failure in our understanding of common language were we to neglect a precise treatment of the structure of knowledge.

Chapter 3

Structures

3.1 General Notion of an Empirical Structure of Knowledge

An empirical structure of knowledge can be formalized in a way that makes clear the procedure for its construction. In general, an empirical structure for knowledge is as follows. First we gather a set K of elements of knowledge, and assign an n-ary relation R^n to all elements of K; giving a set of relations $R^n(K)$. Now, to obtain a representation of K, we will need a function that projects the relations in $R^n(K)$ onto a diagram D; call this function $P()$ and call the resulting $P(R^n(K))$ a representation of K.

Now, it is clear from this generalization that there are many different choices of R^n and P relative to which we can obtain a representation of a given K. One of the methods I have employed took the trinary relation of Distance $D_{ki,kj,di}$ (see § 5.5 Distance Metrics) as R^n, and some algorithm (NJ, UPGMA, or Bio-NJ; see § 5.10 Visualization, § 6 Results) or piece of phylogenetics software for mapping distances onto trees as a function P. But there are still plenty of different distance relations we might choose, and at that, plenty of ways to transform these relations after they have been obtained (see § 5.9 Statistical Evaluation of Structures). Indeed, the choice of P is also a choice of a type of diagram D, and some choices of P will give different types of diagrams depending on R^n (e.g. using SplitsTree to generate a phylogenetic splits-net can give diagrams that are trees, more or less tree-like, or those that more closely resemble networks). Nonetheless, if K is empirical data, R^n depends on K, and P depends on R^n, then the resulting representation of K will be empirical in the most straightforward way.

At this level of generality we can state some facets of the notion of conceptual relatedness more exactly (See § 6.). But I believe the most striking advantage is the separation of the method for generating structures into three discrete parts (See § 5.1 Notes on General Method).

3.2 The Meaning of Structural Representations of Domains of Knowledge

The possible meanings of structural representations are more varied than the representations themselves. Some seek to capture notions of differences in the scale of theoretical entities, others reflect the organization of departments in academic institutions—the historical splitting of disciplines into subdisciplines—and others still (of paramount interest to me) attempt to model the relations between the conceptual apparatuses of distinct disciplines.

The significance of structural representations is a function of how they are used. Surely some people want them to be ontological—based on the ontological commitments of theories—and thus representations that capture differences in scale (or possible theoretical reductions) are put forth. Likewise, some want the structure to indicate the interdisciplinary relationships that obtain between sciences and thus prefer networks that graph the citations between major journals (such as EigenFactor). I am interested in addressing pre-existing philosophical intuitions about conceptual structures. All of these representational goals structure our understanding of scientific knowledge, yet to entirely different ends. So it would be a fool's errand to analyse these structures, offered for different uses, as if they were aiming at the same truth, mutually consistent, or subject to the same critique.

Put another way, *structural representations of knowledge are model dependent and interest relative.* And while it is just these features of structural representations that make them useful, they also open the door widely to equivocation: one would be rather disappointed to find out that, far from supporting any sort of hierarchy of consensus or complexity, the work of Fanelli and Glänzel (2013) supports only a hierarchy of writing practices[1]!

There is a reciprocal relationship between our understanding of domains of knowledge and the patters of structural representation. Take the following as an example. One could set out to find the most appropriate hierarchical relationship to represent the differences in scale amongst the natural sciences, only to find that the disciplines

[1] While this might be an all too literal interpretation of the HOS according to Fanelli and Glänzel (2013), one can hardly be blamed for taking a deflationary and excessively flat stance in an area that is such a mix of odd justifications for philosophical ideas and untutored lay intuitions of scientific practice.

often employ entities from a diverse range of scales—thus necessitating a branched hierarchy. Similarly, one could work backwards from a set of theoretical entities arranged hierarchically by scale to a set of gerrymandered disciplines that mention only those entities within a specified range—thus necessitating a radical divergence from more traditional disciplinary boundaries. In general this is just a consequence of the ability to both fit data to a structure and reciprocally to simulate data to fit a predefined structure (it is common practice within phylogenetics to both fit phylogenetic trees to data, and to simulate data to fit a predefined phylogenetic tree). This reciprocal relationship affects not only the modelling of structures but also our folk-philosophy of them. It is easy to project some naïve structure onto the whole of science without first examining the status quo of science to see if the fit is Procrustean. It is exactly this kind of false prioritization of structures over data (Fanelli and Glänzel 2013; Buckner et al. 2011) that I see fit to criticize amongst philosophers who have constructed structures by consulting their intuitions (Compt 1835).

Chapter 4

Data Sources

4.1 Wikipedia

The most common objection to Wikipedia as an academic reference is its supposed inaccuracy. Wikipedia is not peer reviewed in the same way that academic journal articles are: Wikipedia has a large group of editors who check, cross-reference, newly made changes to articles, while major journals tend to require review and acceptance by a small group of professionals within the discipline of the article[1]. Unlike peer-reviewed articles, which are subjected to an intensive editing process by their committed authors before final publication, Wikipedia articles are in a state of flux of partial responsibility; pages are constantly being added to, edited, and even vandalized. Certainly these are the prices paid for being public.

It is important not to run two issues together: Wikipedia's legitimacy as an academic reference and its use as a research tool (or source of personal background information). Indeed, if we prefer peer-review, we do not need to take Wikipedia's word for anything—there is a large, and growing, quantity of peer-reviewed scholarship that has taken up the task of analyzing Wikipedia.

A debate between Encyclopædia Britannica (EB) and Nature highlighted the extent to which the legitimacy of Wikipedia has become a matter of scholarly debate. Nature claimed that Wikipedia and EB have unequal yet comparable numbers of errors in their entries about scientific topics (Giles J. 2005), while EB claimed that Nature's study was "Fundamentally Flawed" (Encyclopædia Britannica, Inc. 2006). To my mind, and for the present purpose, the average number of errors present in Wikipedia articles is irrelevant. Indeed, whether or not Wikipedia is frequently in error, it still represents the view of a large number of contributing authors (at least

[1] This is not to say that just being a journal puts a resource necessarily on better footing; examples abound of predatory publishing practices are an obvious case where journal legitimacy can be taken falsely for granted.

10,000 “Wikipedian” editors with more than 14,000 edits each (as of January, 2016)); and large groups are expected to make errors.

Instead of the number of errors on Wikipedia one might instead be concerned with the relationship between articles and sources. Finn Årup Nielsen (2008) published a report on the number of citations of journal articles on scientific Wikipedia entries. Nielsen concludes that,

> An increasing use of structured citation markup and good agreement with the citation pattern seen in the scientific literature though with a slight tendency to cite articles in high-impact journals such as Nature and Science. These results increase confidence in Wikipedia as a good information organizer for science in general. (Nielsen 2008)

And even more optimistically,

> [U]se of structured scientific citations in Wikipedia will very likely continue to grow and increasingly benefit researchers that look for well-organized pointers to original research. [ibid]

Today, seven years later, we can hope that this trend has continued without ebbing significantly.

4.2 Stanford Encyclopedia of Philosophy

SEP perhaps occupies a milieu between the justificatory status of Wikipedia and peer reviewed articles. Certainly in the domain of encyclopedias of philosophy, given the scope and constant updating of articles, SEP is the ultimate (Buckner et al. 2011). While the format for SEP articles is undeniably encyclopedic, the articles are commissioned from professionals in the field, and reviewed by a competent editorial board and peers (*The Stanford Encyclopedia of Philosophy*, Editorial Practices).

Compared to most Wikipedia articles on philosophical topics, SEP articles are quite long and, one could argue (from more than authority), more detailed. Nonetheless, while Wikipedia articles are hyperlinked and SEP articles possess only a small number of “related topics”, this poses no significant problem for the analysis of SEP (see § 5.2-5.4, 6.), since parsing and filtering methods can be used that extract a more significant stock of keywords from the body of the article.

Prior work has been done on mapping SEP by the Indiana Philosophy Ontology (InPhO) project, directed by Colin Allen at the University of Indiana. As well as some Latent Dirichlet Allocation (LDA) based topic modelling that is beyond the scope of this thesis, the InPhO also offers a "taxonomy" or "computational ontology" of philosophy. Buckner et al. (2011), members of InPhO, characterize the ontology as follows,

> '[C]omputational ontology' denotes a formally-encoded specification of the concepts relevant to a subject domain (including their properties and relations between them) and a hierarchical classification of those concepts into categories and subcategories. (Buckner et al. 2011)

And indeed their taxonomy is hierarchical: a perfect dichotomously branching tree. And while some of the methods involved in generating this taxonomy are certainly more advanced than the simple set theoretic analyses presented here (see § 5.5), they lack the open-ended potential to generate alternative structures enjoyed by phylogenetic (network) analyses (see § 1.6, 3.1). Indeed, their methods suffer from the same problem as Fanelli and Glänzel (2013)—they fail to consider the range of possible structural representations of knowledge—yet at one higher level: they include trees but neglect reticulated trees.

4.3 Encyclopedia as Approximations of Disciplinary Discourse

It is obvious that encyclopedic articles pale in breadth when compared to the swaths of literature available from journals, books and databases. Nonetheless an encyclopaedia has two features that lend themselves specifically to the study of the organization of knowledge at the level of disciplines: labelling and connection.

When attempting to find a way to classify texts, categorize them, often a major obstacle is finding sources of text that are already labelled with a given classification. For example, if we sought to classify a set of ancient texts by author, one way to begin this analysis would be to obtain a set of documents of known authorship. Now, my aim here is not particularly the classification of discourses into disciplinary categories—not to say that the methods employed here might not be used to that end—but examination of the relationships between already classified discourses. Wikipedia authors are engaged in active and often heated debate about how to categorize and organize their pages, titles and subsections, and authors of SEP are commissioned to write on a specialized topic. This puts Wikipedia and SEP pages in the special position of being entirely labeled discourses. And while many journals do add categories to their search function, or have users label pages by keyword (notably, PLOS ONE), these categories are either sparse or not necessarily amenable to disciplinary organization; e.g. on PLOS ONE 'bacterial genetics' is a subject area, but so is 'biomarkers', something I would perhaps call a "topic" but not a discipline[2].

Journal and encyclopaedia articles are both structured to include connections between them: citations in the case of journal articles, and often hyperlinked text in online encyclopaedia (see § 1.4, 1.5). In much the same way that citation analysis can infer relationships from connections drawn by citations between articles, hyperlinks serve as evidence of a conceptual connection between hyperlinked discourses. But the conceptual waters are muddy in both camps. Journals articles in different disciplines often have different citation practices[3], some authors cite others merely out of respect (politics) or lack thereof, and even experienced authors do not cite all relevant work. Similarly, Wikipedia pages are hyperlinked to disproportionate degrees by different authors, similar political linking practices exist, and often many irrelevant things are linked merely for completeness.

[2] Certainly the reasoning for classifying academic journal articles by publication is pragmatic: researchers do not go out in search of all the articles within a discipline but instead seek specific sources to advance their research or justify it. So disciplinary organization is not, nor need it be, sufficiently fine grained in those contexts.

[3] This fact is one of the things that Fanelli and Glänzel (2013) used as a means of making structural distinctions between disciplines.

Chapter 5

Methods

5.1 Notes on General Method

Philosophers and philologists have traditionally studied the relationships between disciplines historically or by consulting their intuitions. Instead of historicism or private scholarly intuitions, my methods attempt to account for public discourse and thus public intuitions. In this respect I share the inclinations of bibliometric analysts (Fanelli and Glänzel 2013), co-word analysts (Qin 1993., Callon, Courtial and Laville 1990), network theorists (Zlatić et al. 2006., Masucci et al. 2011), and the use of statistical assessments of publicly generated data can be considered in the spirit of experimental philosophy.

The general notion of a representation of a structure of knowledge (developed in § 3.1) can be transferred to the structure of disciplines when the "elements of knowledge" analysed are taken to be representative of disciplines. This process can be broken down into three parts.

Gathering Disciplinary Discourses

The problem of choosing which discourses were to count as disciplines was tackled in three ways: 1) by defaulting to the opinion of the Wikipedian editors and simply gathering disciplines from curated lists, 2) by choosing small canonical sets that I believe everyone agrees are genuine disciplines then expanding on these in turn, and 3) by algorithmically gathering pages that were close to a given page in the network of wikipedia, then filtering these pages based on common disciplinary markers like ending in 'omics' or 'logy'. (See § 7, Ontology).

I began by gathering a corpus of discourses for a set of disciplines and extracting keywords from the text—relevant text being that which is not expected to appear in all sources. Keywords obtained from hyperlinks should be expected to contain a variety of theoretical terms relevant to the discipline in question, but also names of famous practitioners and commonly employed resources. So the elements of the set of knowledge K were always broken down into lists of keywords and associated disciplinary titles.

Defining Relationships Between Discourses

I employed a set of distance metrics, to give an approximation of the difference between all pairs of discourses. Not all relations of distance are equal in the eyes of human intuition, as I show below (§ 6.), many distance metrics gave wildly unintuitive results, or latched onto features of keywords that were wholly irrelevant to their conceptual relatedness.

The ternary relation of distance $D_{ki,kj,di}$ was not the only relation, defined on the set of discourses, used in representations; keyword co-occurrence relations were also investigated. Graphs were constructed where vertices were keywords or disciplines, and edges were drawn whenever disciplines shared particular keywords. This basic ternary relation[1] in turn allowed the definition of relations of even higher-arity (i.e. R^n) on the set of disciplines, such as the computation of various kinds of graph modularity (§ 5.6-5.7).

Mapping Relationships onto Diagrams

Distance data was mapped onto simple linear hierarchies by extracting distance matrix columns and ordering each by distance from 0. Distance information was also passed though software (SplitsTree) that constructs phylogenetic trees and networks, using standard tree and network building algorithms (NJ, UPGMA, Bio-NJ). This allowed a visualization of the conceptual relatedness between all members of the corpus of disciplinary discourses, comparison of distance metrics, and of local branching and global domain architecture.

Similarly, domain structure computed on word co-occurrence graphs was mapped onto trees to allow comparison of modularity and domain architecture. Co-occurrence graphs and modularity results were themselves represented on hive-plots (§ 5.10).

[1] "discipline i and j share term x", or $\{x \in C_i \wedge x \in C_j\}$ where C_i is the set of concepts for discipline i

5.2 Web Parsing

Data was algorithmically gathered from Wikipedia pages chosen from the set pages on a Wikipedia-official list of academic disciplines. Early analyses were conducted by parsing out the “list of academic disciplines and subdisciplines”, extracting subsections of the disciplines, their subdisciplines, and gathering page data from them. This method was in a way too much and too little: while it often gave discipline sets that were too large to be visually represented with any clarity, it also neglected some important subdisciplines (such as “functional genomics”), so more hand-picked discipline choosing methods were preferred (see § 5.4).

Parsing methods were easily transferable to non-Wikipedia sources like SEP, with the minor exception that SEP does not contain hyperlinks in the body of the text. Instead SEP has a series of “related topics” listed at the bottom of each page, so any link to a “topic” was treated as roughly analogous to page for a philosophical discipline[2].

5.3 Keyword Extraction

The hyperlink for each discipline included in the Wikipedia analysis was followed (”crawled”) and each link on the page was gathered, stripped of superfluous characters, and saved as a keyword associated with the discipline in question. Using hyperlinks as keywords has three advantages: they are *Unique*, *Univocal* and *Ubiquitous*. Respectively, there is usually only one hyperlink of a given name per page, they always lead to exactly one page and never to two or more, and there are usually plenty of them to be used in the analysis.

Yet no matter how well hyperlinked a page is, there will always be words that deserve a hyperlink that do not receive one. To catch these words, and in hopes of extracting more *relevant* keywords, I also employed a full text analysis. Full text from pages was extracted using BeautifulSoup4 (©1996-2016 Leonard Richardson), and was tokenized and had stop words removed using NLTK (Bird et al. 2009). The

[2] Although this fragmentation of SEP into many philosophical subdisciplines is perhaps philosophically suspect, there is no resource managed through SEP to collect and demarcate “official” subdisciplines. But, moreover, most entries on SEP can be considered sufficiently general to enjoy the same *sui generis* status as articles on subdisciplines of, say, biology. I cannot imagine a very convincing argument that ‘Oology’ is a discipline while ‘Animalism’ is not.

process of tokenization and filtering were similar when academic journal articles were used, except the PDF parser PDF2TEXT was used to convert articles to plain text format.

5.4 Choosing Datasets

To avoid the representational burden of large collections of disciplines—some truly uninteresting—I have restricted many analyses to data subsets: the FEGC and FEG datasets (see Methods § 5.8), and three canonical sets (two slightly more expanded, see Table 5.1). When an alternative dataset is employed this will be explicit. The Canonical and Canonical+ datasets were chosen by hand to reflect a reasonable distribution of disciplines, while the Canonical++ dataset was gathered algorithmically from a page containing an outline of the disciplines covered on Wikipedia (Available at Wikipedia: Outline of Disciplines, Xs indicate absent disciplines). The FEG dataset was gathered by crawling the pages for 'Functional_Genomics', 'Genomics' and 'Evolutionary_Biology' and extracting potential disciplines, whereas in FEGC 'Chemistry' was added as an expected outlier.

Any data set could be used in theory, as long as there is a Wikipedia, SEP page, or a convertible PDF available. In fact, any combination of these data sources could be used together if desired.

5.5 Distance Metrics

The distance metric employed was a 'harmonic mean of conceptual divergence'—an average of the magnitudes of keyword set differences. That two disciplines diverge from each other is perhaps a misnomer, since we might falsely interpret this to mean they were once quite similar and have since grown different, while this set theoretic metric is ahistorical. Nonetheless the notion of divergence does capture the fact that different encyclopaedic articles on disciplines often show some variable and characteristic degrees of overlap and distinction. *This is precisely the sense in which the absence of a strictly inclusive or exclusive hierarchy allows the construction of representations* (§ 1.6).

The conceptual divergence between each discipline U_{ij} was calculated, as below,

Table 5.1: Canonical, Canonical+ and (truncated) Canonical++ datasets.

Canonical	Canonical+	Canonical++
Sociology	Sociology	Sociology
Psychology	Psychology	Psychology
Philosophy	Philosophy	Philosophy
Anthropology	Anthropology	Anthropology
Physics	Physics	Physics
Biology	Biology	Biology
Cosmology	Cosmology	XXXXXXX
Quantum_mechanics	Quantum_mechanics	Quantum_mechanics
Chemistry	Chemistry	Chemistry
Geology	Geology	XXXXXXX
Biochemistry	Biochemistry	Biochemistry
	Logic	Logic
	Mathematics	Mathematics
	Meteorology	Meteorology
		Humanities
		History
		Linguistics
		Visual_arts
		Religious_studies
		Cultural_studies
		Economics
		Gender_and_sex_studies
		Journalism
		Computer_science
		...

with C_j being the set of concepts gathered as hyperlinks from discipline j.

$$U_{ij} = \{x \mid x \in C_i \land x \notin C_j\} \tag{5.1}$$

The ratio X_{ij} of the size of the conceptual divergence to that of the initial concept set was taken, essentially giving a measure of the proportion of C_i that makes up the difference with C_j,

$$X_{ij} = \frac{|U_{ij}|}{|C_i|} \tag{5.2}$$

Finally, the he harmonic mean H_{ij} was calculated for each discipline pair as below. This gave a symmetric measure of distance, i.e. one where $H_{ij} = H_{ji}$

$$H_{ij} = \frac{2 \cdot X_{ij} \cdot X_{ji}}{X_{ij} + X_{ji}} \tag{5.3}$$

5.6 Network Construction

Networks were constructed using iGraph (Csardi and Nepusz 2006) and followed two closely related methods—distinguished by what was to count as a node in the graph. In the first type (G_c), nodes included both keywords and the name of each page from which keywords were gathered (disciplines). Nodes for keywords were iteratively connected first to the page on which they were gathered, then to any other page on which they appeared—thus disciplines were connected via their keywords. In the second type (G_t), nodes included only the names of each page (disciplines) and nodes were connected whenever they shared a keyword—thus disciplines were directly connected. For each keyword shared between disciplines i and j in a G_t the weight of the edge e_{ij} was increased by one.

5.7 Network Evaluation

For smaller graphs like G_t when less than 100 disciplines were included, the optimal modularity method offered by iGraph could be used. But larger graphs like G_c or any G_t with more than 100 nodes would exceed time constraints, so weaker but sufficient methods such as the "Community Edge Betweenness" or "Fast-Greedy" clustering algorithms were used (Girvan and Newman 2002).

Modules were mapped onto trees as follows. Once tree image files were converted into SVG format, the leaves whose name corresponds to each member of a module were identified and coloured according to a randomly selected colour palate. The coloured trees and splits networks were then statistically assessed for the similarity of their domain architecture and modularity.

See § 5.7 (Statistical Evaluation of Keyword Datasets) below for details on how networks were used to statistically evaluate the importance of specific keywords, also see § 5.9 (Statistical Evaluation of Structures).

5.8 Statistical Evaluation of Keyword Datasets

Keywords were stored in lists associated with the page on which they were gathered. This allows a barrage of standard statistical measures to be applied to the sizes of these lists—ex. 'Functional genomics' has 56 hyperlinked keywords while

‘Evolutionary biology’ has 167. Thus as with any standard numeric dataset one can calculate the mean, median, standard deviation, variance, quartiles, maximum and minimum, and compare these values between datasets—as a measure of the quality of data gathered or differences in the effectiveness of crawling methods.

Yet the nature of this analysis allows some non-standard measures to be applied. For example, when looking at an entire dataset of keywords, it is of interest to examine which keywords from each list are shared between lists: the informative subset. A keyword t_i from a list of keywords C_i is informative for discipline i just in case t_i is also contained in some C_j for the same dataset (where $j \neq i$). For the FEGC dataset, 41 of the keywords of ‘Functional genomics’ were informative while 141 were for ‘Evolutionary biology’. This leads to a set of related statistical notions such as the informative average, informative standard deviation, informative variance, informative fraction, etc.

Since the informative elements of each list are relative to the elements of every other list in the dataset—to be informative a term must be shared with at least one other page—the informative fraction (inffr(d_i)) of keywords for a given page (d_i) will change depending on the method of collection. An informative fraction close to 1 indicates that the page was crawled along with other closely related pages, while an informative fraction close to zero indicates little relation to other pages in the dataset—ex. for the FEGC dataset, inffr(‘Functional genomics’) = 0.732, inffr(‘Evolutionary biology’) = 0.844, inffr(‘Genomics’) = 0.690, while inffr(‘Chemistry’) is only 0.229. When ‘Chemistry’ is left out of the crawling dataset (the FEG dataset), inffr(‘Genomics’) increases to 0.694, and when ‘Biochemistry’ is added to the dataset (BFEGC) there is an increase to inffr(‘Chemistry’) = 0.280.

The construction of networks also allowed statistical evaluation of keyword datasets. After a graph is constructed (G_c or G_t), it is possible to compute the betweenness centrality, $g(v_i)$, of each node (v)—a measure of the number of shortest paths from every node to every other that passes through that node. We can think of $g(v_i)$ as a measure of the importance of v in connecting G_x. $g(v_i)$ values are easier to interpret when normalized in the following way,

$$g_{normalized}(v_i) = \frac{g(v_i) - \min G(v)}{\max G(v) - \min G(v)} \tag{5.4}$$

Where G(v) is the set of all $g(v_i)$, this procedure ensures that $g_{normalized}(v_i) \in [0...1]$.

Table 5.2: $g(v_i)$ values for sample of FEGC and FEG datasets

	Datasets	
Disciplines	FEGC	FEG
Functional Genomics	0.1387	0.1485
Evolutionary Biology	0.6314	0.6733
Genomics	0.9665	0.9934
Chemistry	0.1437	NA

For example, Table 5.2 shows the differences in $g_{normalized}(v_i)$ (hereafter just $g(v_i)$) after the removal of 'Chemistry' form the FEGC dataset. The betweenness of each major node increased after the removal of 'Chemistry', i.e. these nodes became relatively more important in connecting the network in the absence of 'Chemistry'.

5.9 Statistical Evaluation of Structures

Evaluation of SplitsTree Structures

In order to generate a SplitsTree structure (tree, reticulated tree, network) from a matrix of distance data is it necessary to skew the values of the matrix[3] so that the final branch length between two taxa of the structure will be different from the value entered into the original matrix. Consider a matrix like the following, where d_{ij} is the distance between disciplines i and j, and $d_{kk} = 0$.

$$\begin{bmatrix} 0 & d_{12} & d_{13} & d_{14} & d_{15} \\ d_{21} & 0 & d_{23} & d_{24} & d_{25} \\ d_{31} & d_{32} & 0 & d_{34} & d_{35} \\ d_{41} & d_{42} & d_{43} & 0 & d_{45} \\ d_{51} & d_{52} & d_{53} & d_{54} & 0 \end{bmatrix}$$

When this matrix is represented, as a tree for example, one can follow the branch lengths within the tree to reconstruct the distance matrix. The recovered distance Δd_{ij} will differ from d_{ij}, giving a matrix as follows.

[3] Unless the matrix is already perfectly tree-like, network-like, etc.

$$\begin{bmatrix} \Delta 0 & \Delta d_{12} & \Delta d_{13} & \Delta d_{14} & \Delta d_{15} \\ \Delta d_{21} & \Delta 0 & \Delta d_{23} & \Delta d_{24} & \Delta d_{25} \\ \Delta d_{31} & \Delta d_{32} & \Delta 0 & \Delta d_{34} & \Delta d_{35} \\ \Delta d_{41} & \Delta d_{42} & \Delta d_{43} & \Delta 0 & \Delta d_{45} \\ \Delta d_{51} & \Delta d_{52} & \Delta d_{53} & \Delta d_{54} & \Delta 0 \end{bmatrix}$$

Since the two matrices will be the same size, one can vectorize both and plot Δd_{ij} Vs d_{ij} to obtain a graph for regression analysis. Vectorization proceeds as follows. A matrix can be represented as a list of lists: each row of the matrix is treated as a list entry with index equal to the ordering of rows—although this is arbitrary, it only matters that one use the same ordering/vectorization for both . After obtaining a list of lists the vector is obtained by transferring the elements to a new list in order. One can then plot the set of ordered pairs $< x_i, y_i >$ where x is from the original matrix and y is from the distorted matrix and i is the index of both list values. If the matrix did not need to be distorted at all, then $x_i = y_i$ for all i ($R^2 = 1$), and if not, then the R^2 value will give an indication of degree of distortion.

A possible problem with measuring the correlation between matrices this way is that it assumes the independence of distances when calculating the p-value of the correlation[4]. A statistical test called the Mantle test (first described in Mantle 1967) accounts for this dependence. So the p-value measured by regression will be lower than that measured by the Mantle test, but the R^2 value itself will be identical (see Figure 10 and Figure 11 for a comparison of both statistics). Nonetheless, the p-values for both results usually far exceed the requirements for a statistically significant test so, when no significant difference will result, only the p-value from regression will be reported.

Evaluation of Linear Hierarchies

Once a distance matrix is obtained it is possible to examine a set of linear orderings. Since the distance matrix specifies the distance between every pair of disciplines, it is possible to extract just the set of distances between every discipline and

[4] An assumption that is obviously violated. To change the distance between x and y would require adding keywords to either x or y, and this could change the distance between them and some other discipline z.

a discipline of interest, a single row of the matrix, then arrange these disciplines in ascending order. Since every column will have one discipline compared to itself, a distance of zero, that discipline becomes a *pole*—a discipline at the bottom or top of the hierarchy.

To evaluate these hierarchies one can reconstruct a distance matrix from only the distances present within the single column of the matrix used to generate the hierarchy: a significant loss in information to be sure. Hierarchies can then be ranked by the R^2 value of their vectorized matrix plotted against the vector of the original matrix (Table 5).

Evaluation of Fit Between Modularity and Structure

Since modularity results are not guaranteed to agree with the domain architecture or distances represented in trees (see Figure 23) a method of assessing the fit between modularity and trees was developed.

Perhaps the simplest method of assessing the fit between structures and a modularity result is counting the number of Sub-tree Pruning and Re-graft (SPR) operations needed to put the members of each module within the same domain—and perhaps relativizing this count to the total number of leaves on the tree to give a distance measure (RSPR). For example, a tree with 10 leaves and 1 leaf not branching from within a domain consisting only of members of its assigned module would have an RSPR = 0.10. Unfortunately, this method is sensitive only to the topology of the tree and not to its metric aspects: SPRs do not account for branch lengths.

One method that is sensitive to branch lengths when assessing module fit is the subtype diversity ratio[5] of a module (SDR_m), which can be defined as the ratio of the mean within-module (within-subtype) pairwise distance to the mean between-module pairwise distance (Rambaut et al. 2001).

$$SDR_m = \frac{\sum_{i,j}^{n} d_{ij}}{\sum_{i,k}^{n} d_{ik}} \cdot \frac{N^*}{N} \text{ : for leaves i, j in module } m \in M \text{ and k not in } m \qquad (5.5)$$

Where N is the number of pairwise combinations of elements of m and N^* is the number of pairwise combinations of elements between m and all other modules[6].

[5] A measure originally developed to assess the fit between HIV viral subtypes and viral phylogenies.

[6] More precisely, $N = |m \times m|$, and $N^* = \sum_k^n |m \times k|$ for $k \in M$ s.t. $k \neq m$.

The SDR_m is a measure of how "tightly" m is clustered relative to the rest of the tree, when SDR_m is low, the module is tight, when it is high it is loose or sparse. In an inversion of the unfortunateness seen in the RSPR method, the SDR method accounts only for the metric aspects of the tree and not for its topology—the SDR of a module will be the same regardless of the branching pattern of the tree, so long as the distances remain the same.

A method was developed that measures the ratio of the Distance Separating Elements of a module (DSE) from a domain to the Total (non-leaf) Intramodule Distance (TID) between internal nodes in a module to account for both the metric and topological aspects of trees when assessing the fit of modularity results. As such, this is a measure of the Topological and Metric Disagreement for a module (TMD_m) of a tree from a modularity result. To my knowledge this method is original.

When s_α^β is defined[7] as the length of the split separating the sets of leaves α and β, TMD_m can be defined as follows,

$$TMD_m = \frac{DSE_m}{TID_m} \tag{5.6}$$

$$DSE_m = \sum s_\alpha^\beta : \text{for } \alpha, \beta \nsubseteq m \text{ and } \alpha \cap m \neq \varnothing \text{ and } \beta \cap m \neq \varnothing \tag{5.7}$$

$$TID_m = \sum s_\gamma^\delta : \text{for } |\gamma| \neq 1 \neq |\delta| \text{ and } \gamma \cap m \neq \varnothing \text{ and } \delta \cap m \neq \varnothing \tag{5.8}$$

I will explain each of the side conditions in turn. Firstly, for the DSE_m, specifying that both α and β must not be subsets of m ensures that one does not sum those splits that are parts of domains consisting entirely of members of a module[8], as these splits do not *separate* elements of the module. That the intersection of the module m and the sets $(\alpha, \beta, \delta, \gamma)$ that each split separates must be non-empty is to specify that one only sum those splits that separate members of the module under consideration—that is, one only sum those splits that are between members of a module. Finally, for TID_m, that $|\gamma|$ and $|\delta|$ cannot equal 1 ensure that no leaf split

[7] For a concrete example, $s_{\{`Biology'\}}^{\{everything\_else\}}$ would be the length of the terminal branch with 'Biology' as a leaf.

[8] For lack of a better term, and in the spirit of the term 'monophyletic' perhaps one could call these domains 'monomoduletic'.

is counted. This last condition is important since a leaf would always disagree with a module if there were more than one element of the module and it was a member of the module.

These conditions together ensure that the TMD_m is a distance measure: when no elements of a module are within the same domain (monophyletic group) $TMD_m = 1$ and when all are "monomoduletic" $TMD_m = 0$. Variation emerges between these extremes that indicate the degree of fit between a modularity result and a tree consisting of elements of that module.

Together, SPR, SDR_m and TMD_m can give us a picture—along various axes of analysis—of the fit between a modularity result and the structures it is mapped onto.

5.10 Visualization

Distance matrices can be fed, once converted to NEXUS format, into software used to generate phylogenetic trees. SplitsTree is one such piece of software that generates phylogenetic networks, or a splits-net, by the method of split decomposition (see: Bandelt and Dress 1992; Huson and Bryant 2006; Dopazo et al. 1993).

SplitsTree can also be used to generate rooted and unrooted bifurcating trees (hierarchical and non-hierarchical trees). To generate trees, SplitsTree has the option of using two different algorithms, each with their own set of underlying assumptions.

UPGMA

The Unweighted Pair Group Method with Arithmetic Mean (UPGMA) algorithm (Sokal and Michener 1958) is a currently unpopular hierarchical clustering method for constructing phylogenetic trees. This is because UPGMA assumes a constant rate of evolution, sequence diverge—an assumption that, presumably, is often violated.

The algorithm functions by first joining the closest elements in the distance matrix and making them sister taxa within the tree with identical branch lengths equal to half their distance. Then secondly calculating a new matrix by taking the mean of the pairwise distance between the joined elements and everything else. So for a distance matrix with disciplines $i...n$ where the distance d_{ij} was the smallest, i and j would be merged into $\bar{ij}$, and the new distance matrix would be calculated as follows,

$$d_{\bar{i}jk} = \frac{d_{ik} + d_{jk}}{2} \tag{5.9}$$

Neighbour Joining

The Neighbour Joining (NJ) algorithm, and its variants (Bio-NJ), all contain the underlying assumption that rates of evolution might differ between sister taxa. This, to my mind, is a much more realistic assumption for both DNA sequence evolution and the differences that emerge between discourses in the process of generating knowledge. There seems to be no reason to assume that since two disciplines are most closely related, that they are just as different from their nearest relative[9].

The mathematical properties of NJ are too complex to reiterate here (see Saitou and Nei 1987). But, similarly to UPGMA, NJ is a bottom up hierarchical classification algorithm that takes a distance matrix, except first modifies it, then chooses the nearest neighbours to join by a new node in a star tree. Then a separate calculation is preformed to decide the branch lengths between now sister taxa, before recalculating the modified distance matrix.

So, excluding linear hierarchies and rhizomes, SplitsTree allows us to visualize the structures outlined in § 1.2-1.4, providing a means of comparison between our intuitions and the encyclopedic data.

D3 Graph Visualizations

Networks were represented using the D3 graph visualization JavaScript library (Botstock et al. 2011). Since the relevant graphs were too large to be meaningfully displaying in a simple network or force directed graph, an approach based on the *hive-plot* of Krzywinski et al. (2012) was used. These plots avoid the problem of computing the optimal layout for a graph (and hoping that clusters appear clustered) by stipulating that clusters of related nodes lie in radial axes, with connections between nodes of different clusters shown as edges between axes and within-cluster connections compressed along the axis (see Figure 27). The size of nodes v_i are represented proportional to $g(v_i)$ and edge widths as $|\{x \mid x \in C_i \wedge x \in C_j\}|$.

Rooting Trees and Networks

Trees and networks were rooted both by hand and automatically by SplitsTree. In

[9] Indeed, this does not seem like the kind of thing about which one would usually even *have* intuitions prior to seeing the results of using algorithmic methods of classification.

UPGMA trees the root was placed at the midpoint of the tree (due to the assumption of equal rates of evolution), while in NJ trees the root was placed next to the last taxon to be resolved in the tree during hierarchical clustering. This taxon will be the one that never had the smallest distance to another taxa during tree construction. Both of these methods derive from assumptions about sequence divergence, and I am very skeptical that they can be justified in any analogous way when the input data is not evolved sequences but encyclopedic texts (but see § 7). It seems the roots serve only to highlight the *possibility* for a polarized process of disciplinary organization: a possibility that I believe will be unrealized most often, but certainly not impossible in specific cases.

Consider also the issue of interpreting internal nodes and splits,

> There is an important difference between phylogenetic trees and more general split networks: Any rooted tree has a direct interpretation in evolutionary terms: the leaves represent taxa and the internal nodes represent speciation events. In a (possibly rooted) split network, the internal nodes do not have such a direct interpretation. Instead, split networks must be viewed on a more abstract level as networks giving a visual representation of incompatible signals, that is, showing how "tree-like" or "certain" parts of a phylogeny are. (Huson and Bryant (2006) *User Manual for SplitsTree4 V4.6*)

Here an analogy with sequence diverge seems more appropriate: in a tree the internal nodes represent major differentiations in conceptual structure, while in a split network the internal structure represents unresolved conceptual signals, or estimates of how certain we can be in differentiations in conceptual structure.

Chapter 6

Results

Wikipedia

Table 5 presents the top 3 results of a linear ordering of the canonical set of disciplines. Although all the best results have p-values $<< 0.05$, the value of the regression of the distorted and original matrices don't show correlation. This was to be expected since constructing a linear ordering in this way significantly reduces the amount of information used from the original distance matrix[1].

Figure 1 depicts a simple tree representation of the distances computed by analyzing Wikipedia pages for the canonical set of disciplines. In ascending order it appears almost exactly to match the intuitive linear hierarchical structures. In fact, appart from cosmology, it appears quite similar to what one might expect from a branching hierarchy of physical reductionism (the scale of the entities in question roughly increases as you move away from the root).

To obtain a result that so closely matches our intuitions about the conceptual organization of disciplines is surprising—indeed, amongst so much philosophical doubt, to obtain a result by empirical means that agrees with a priori philosophical speculation is always quite surprising. But this image is only part of the picture: it was generated using a particular set of input disciplines, a particular algorithm (namely UPGMA, see Sokal and Michener 1958), and parameters (tree must be rooted, unweighted). So it would be unfair and unrealistic to insist that this tree (these methods) be given special status *just* on the grounds that it (they) agrees with some common intuitions; especially since the R^2 value (0.477) is so low compared to other trees given the same dataset. Indeed, subtle changes in the choices of parameters here can give widely different trees.

[1] Certainly the methods of linearizing used by Fanelli and Glänzel (2013) were more advanced than the simple matrix manipulations presented here. Nonetheless, since the linear representations were constructed with the same distance matrices used in (reticulated) trees, they should be considered as the proper point of departure.

For example, Figure 2 was generated using the same dataset, algorithm, and having the same topology, but allowing the use of weights when computing the tree. The difference in branch lengths—most prominent in the case of Geology—indicates a greater difference in linked terminology than made apparent by the previous tree, and likewise the close proximity of quantum mechanics and general physics was underrepresented in the previous tree[2].

In Figure 3 the Neighbour-Joining algorithm was used, and it is once again unweighted. The bifurcation between the branch containing biology and that containing philosophy is significantly more pronounced than in either Figure 1 or Figure 2. Figure 3 is hardly in agreement with the expectations of physical reductionism—at least, significantly less so than Figure 1. This is not a mark against it. Indeed, this kind of major bifurcation in the tree seems to be what Dupré had in mind when he suggested branching to accommodate the split between micro- and macrophysics (See § 1.2, Tree-Like Hierarchy). It still does accord with a quite natural division of disciplines into physical, life, and social sciences, although viewing the tree in rooted form perhaps obscures this interpretation.

Examining the unrooted version in Figure 4, one can clearly see three branches with intuitively physical, life and social science leaves (again, excluding geology, which here as elsewhere appears to be the nonconformist). Amongst those with some familiarity in looking at phylogenetic trees, there is an immediate temptation to interpret something more than clustering into rooted trees, i.e. polarized relations of ancestry and descent. The use of an unrooted tree makes more sense when there is no intention of displaying historical information; Figure 4 and its kind are more obviously being used to represent clustering of similarity, distance, amongst the data represented by the leaf disciplines.

While in Figures 1-3 the root was chosen automatically, the root could also be chosen manually. The root will automatically appear adjacent to the most distant leaf, the outgroup, so one can choose the location of the root by choosing the outgroup. Taking the same parameters as given for Figure 3 and manually setting the outgroup to philosophy places the root of the tree adjacent to this leaf—consequentially producing the structural arrangement of the canonical disciplines seen in Figure 5.

[2] R^2 values for both trees are the same since they are calculated from the weighted version of the tree.

Setting aside intentional parameterizations overemphasizing the importance of philosophy, that this kind of tree manipulation is possible in general has interesting parallels in how we construct trees that seems intuitive to us. It is always possible, if we do not like the hierarchical arrangement that presents itself, to *turn the tree on its head* and produce the contrary hierarchy (as often happens when people stumbling over the linear-hierarchical model begin to ask whether, with respect to mathematics perhaps, the hierarchy starts or ends there). Nonetheless, this structural manipulability should not be taken with much pessimism: in order to avoid arbitrariness one need only be reasonably explicit about the parameters used and modifications made.

As well as being able to manipulate the outgroups of existing trees one can add leaves to existing trees—almost ad infinitum. Figure 6 was generated for the FEGC dataset mentioned above (Methods § 5.4); it is a cladogram drawn in a slanted fashion for readability. But before we examine large-scale representations, let us consider the slightly less expanded set of disciplines in Figure 7 and Figure 8.

Figure 8 was produced using the same experimental set up as in Figure 7 (a phylogram, with outgroup chosen automatically, generated using the UPGMA algorithm), yet uses the Canonical+ dataset (it has additional leaves for 'Logic', 'Mathematics', 'Archaeology' and 'Meteorology'). Interestingly, the addition of both philosophy and logic to the mix resulted in the tree being automatically rooted closer to philosophy—possibly due to the close similarity of philosophy and logic and their mutually significant difference from most other disciplines. Also, the close grouping of archaeology and anthropology seems to make intuitive sense.

Unfortunately, neither of these UPGMA trees are well supported ($R^2 = 0.477$ and $R^2 = 0.355$ respectively) when compared to the same setup when the Neighbour-Joining algorithm was used[3] (Figure 8-1, $R^2 = 0.771$). Figure 9 also shows a, to my mind more intuitive, grouping of anthropology and archaeology with sociology and psychology, as well as a more obvious natural science domain and an deeper branching physical sciences domain.

Consider the Canonical+ set represented as a rooted tree (Unweighted, NJ algorithm) generated using hyperlinked keyword data in Figure 10, compared to when distance was measured relative to the set of keywords gathered from the entire

[3] Indeed, both have far higher R^2 values than *any* linear ordering

Wikipedia page in Figure 11. Notable differences between Figure 10 and Figure 11 include that geology and meteorology are now sisters, the domain containing physics and the one containing biochemistry are now sisters, the domain containing psychology has moved lower in the tree, and logic has remained the outgroup.

As perhaps is to be expected, the total number of non-trivial (non-stop) words on a given Wikipedia page shows a significant increase compared to just the set of hyperlinked keywords. Table 2 summarizes the differences between the number of keywords in the Canonical+ dataset. On merely the grounds that more keywords ought to provide a better picture of the relationships between discourse, we have reason to believe that Figure 11 is a better tree, approximation, of the structure of discursive relationships than Figure 10.

While Table 2 summarized the absolute differences between the number of keywords associated with each discipline in the Canonical+ set, we can also examine the fraction of these keywords that are shared between members of the set in Table 3, i.e. the informative fraction *inffr(i)* for $i \in Canonical+$. In all cases, for each discipline, the *inffr* increased when whole page text was extracted and processed to obtain keywords. The most significant increase in *inffr* can be seen in 'Geology' and 'Meteorology', perhaps giving us greater confidence in Geology being a sister discipline of Meteorology (Figure 11) than in its branching at the base of the clade containing Anthropology and Archaeology (Figure 10). On average, there was a 40% increase in the fraction of informative terms in each keyword dataset.

As well as wanting to know how many informative keywords are contained in each discipline, we should also know how important a given discipline is in the set. A good proxy for importance is betweenness centrality; $g(v)$ for discipline v in a graph G_x (A G_t graph is used here, see § 5.8 Statistical Evaluation of Keyword Datasets). Table 4 presents the $g(v)$ values for the Canonical+ dataset when hyperlinks are used, and well when full text is extracted. Biology, Chemistry, and Physics indeed obtain the highest $g(v)$ values—are most important in connecting the network—while Philosophy, Cosmology, Logic, Biochemistry, and Sociology all have $g(v) = 0$. Interestingly, when full text is used, $g(v) = 0$ for all $v \in G_t^{wholepage}$. This could only happen if every discipline node was connected to every other node directly[4],

[4] This result obtains even when terms in the intersection of all discipline keyword sets are excluded from the analysis, i.e. everything is connected to everything else, but not because every keyword

i.e. $G_t^{wholepage}$ is a rhizome (see § 1.5, Rhizomes). The rhizomatic character of $G_t^{wholepage}$ makes it impossible to draw distinctions between disciplines based on $g(v)$ values since the graph is unweighted. Nonetheless, by all other metrics considered (keyword density, *inffr*) the underlying dataset for $G_t^{wholepage}$ is better than a mere hyperlink based analysis.

Once we are confident that we have obtained a structural representation worth supporting, we can begin to treat it as data in its own right. Say we are debating whether biochemistry is more biological, or more chemical. Indeed, this debate could go on for some time, since the question is not precisely formed (e.g. it is unclear whether we mean to debate the theoretical entities involved, or the similarity of methods). If the question is then recast as one about the proximity within some structural representation of a knowledge base, we can precisely specify the factors involved in determining this proximity (See § 3.1 General Notion of an Empirical Structure of Knowledge). In this case K, the knowledge base, is the $Canonical+$ data set, the relation R^n is the set theoretic distance metric (described in § 5.5 Distance Metrics), and $P()$ is the Neighbour-Joining algorithm in conjunction with the software SplitsTree. Given this representation of K we can say, from looking at the resulting diagram Figure 11, that biochemistry is indeed closer to biology than to chemistry; they are sisters, and this suffices when we are not concerned with branch lengths. Indeed, while this might not satisfy our desire for a more deeply philosophical or historical answer to the question of the relationship between these disciplines, it certainly enjoys an exactness not possible from oral debates on methodology or theoretical entities.

While the tree in Figure 11 was preferred to that in Figure 10 for reasons pertaining to the keyword dataset used in its construction, both are trees and are limited by the constraints imposed by tree topology. It might be expected *on principle* that a network, like those depicted in Figures 14/15, would be better representations of the hyperlink distance data. This was not observed[5]. The R^2 values for for the Canonical+ Full Text dataset (Figure 11, $R^2 = 0.881$), hyperlink splits network (Figure

set has the same term in it.

[5] This result was so surprising that it was necessary to confirm the method was working on artificially non-treelike data. This gave $R^2_{network} = 1.0$ while $R^2_{tree} = 0.25$, confirming that indeed network-like data would give a higher R^2 when represented as a network

15, $R^2 = 0.877$) are not significantly different[6]. This indicates that the Canonical+ hyperlink keyword dataset is not sufficiently network-like to reject the tree representation.

In contrast, when the full text from each Wikipedia page is used as a keyword dataset ("Whole Page Parsing"), networks indeed show a better R^2 value. The Canonical+ full text dataset (Figure 11), the rooted splits network (Figure 18), and unrooted splits network (Figure 19) *are significantly different*[7] ($R^2 = 0.886, 0.936, 0.936$ respectively). So full text extraction generates distance matrices that are significantly more network like than hyperlink data alone[8].

SEP

SEP does not have in-text hyperlinks—instead employing an often rather small set of hyperlinked "Related Topics" at the bottom of each article. Because of this it was often impossible to obtain any meaningful distance measure from hyperlinks alone, necessitating an extraction and analysis of non-hyperlinked keywords from the text body.

Consider Figure 12. The set of articles included in figure 12 were chosen by hand to reflect a distribution of current philosophy research areas, with an emphasis on philosophy of biology. This star-tree tells us almost nothing about the relationships between this set of disciplines, and certainly nothing about the differences between those articles that all branch from the internal node. It is possible to analyse the same pages by extracting full-text from the body of the article and parsing it into keywords instead of relying merely on the related topics section. Figure 13 is the product of such a full-text analysis—an analysis that clearly shows more distinctions than was accomplished using hyperlinks alone[9]. Of note are the sisterhood of both

[6] Using a $\Delta R^2 \geq 0.01$ as cutoff for significance.

[7] Rooted and unrooted networks will always have the same R^2 value, since rooting does not distort the underlying splits network.

[8] Of course, the set of hyperlinks on a page is a subset of the total set of non-stop words on a page, so the full text extraction does incorperate all the information present in that of the hyperlinks, and some extra.

[9] Statistical comparison of these two trees is deceiving. While Figure 12 does indeed have a higher R^2 than Figure 13, this is because the majority of distances in the matrix generating Figure 11 are 1.0, i.e. $U_{ij} = \emptyset$. So, Figure 12 is certainly a better representation of its distance matrix, but it is a better representation of a far worse dataset.

ethics disciplines, the sisterhood of modal logic and philosophy of mathematics, and, to me, the automatic rooting with the philosophy of biology as outgroup.

Journal Articles: Two Cases Close to Home

The analyses above can be applied to much more local sources of discourse. Figure 20 and Figure 21 were constructed from well cited journal articles written by members of the Comparative Genomics and Evolutionary Bioinformatics group (CGEB), while Figure 22 and Figure 23 were constructed from a selection of publications from full time faculty of the Dalhousie Philosophy Department (DPD) and, for outliers some recent work by my supervisors and colleagues in the Philosophy of Genomics (i.e. Doolittle, Mariscal, Booth and Brunet).

Figure 21 has a higher R^2 value than Figure 20, indicating, as now seems typical for larger and thus more complex datasets, that the network representation is preferred in this case. In either representation one can see a domain corresponding to most of those members of CGEB that are more closely associated with bioinformatics than genomics (namely the domain top right containing 'Blouin' and 'Beiko'). But in Figure 21 the domain is now, more intuitively, represented to exclude 'Harding', a member of CGEB who is arguably less affiliated with bioinformatics and more with genomics. Moreover, 'Doolittle' is ever the outlier.

The network representation of the DPD is again preferred. Some faculty members cluster better with themselves than with others ('Abramson', 'Macintosh', 'Schotch', 'Jeffers', 'Sherwin' and 'Borgerson'), while others have their articles split between different domains of the tree ('Meynell', 'Vinci', 'Campbell' and 'Hymers'). Indeed, this result is probably as affected by sample size and bias than the (in)constancy of each individual's writing, but it does provide some intuitive confidence in the domain-level classifications that we can't obtain from the tree-level R^2 statistic. The domain containing 'Borgerson' and 'Sherwin' is of note, since it seems to correspond well to those authors using bioethical terminology. In Figure 23 this "Bioethics" domain shows a high degree of connection to the "Philosophy of Genomics" papers ('BrunetDoolittle', 'MariscalDoolittle' and 'BoothDoolittle') and those of 'Meynell' and 'Campbell', perhaps attributable to the use of biological terminology in each.

Finally, the analysis of the DPD clearly shows the distinct representational capacities of trees and networks. Bottom right in Figure 22 the paper titled 'Meynell_1' is represented as if it were an ancestor[10] of ‘Campbell_2’, a clearly impossible state of affairs under most interpretations of ancestry and normal writing processes (not to say manuscripts do not occasionally form a kind of revision-lineage). This paradoxical situation disappears once we ascend the representational hierarchy to a structure (network) with sufficient capacity to represent small differences. Figure 22 shows both papers as closely related yet implies no such absurd relations of ancestry.

Mapping Modularity

Figure 24 presents a mapping the results of optimal modularity on a G_t graph onto a tree (NJ, Whole Page Parsing, Weighted). The tree shows 3 out of 5 sister taxa as falling within the same cluster for clusters m_0 and m_1, as well as Logic-Philosophy-Mathematics and Physics-Quantum_Mechanics-Cosmology domains. Unexpectedly, the cluster containing Biology and Biochemistry does not contain Chemistry and does include Anthropology and Archaeology.

The tree requires only a few subtree pruning and regraft operations to restore complete module-domain isometry, having an $RSPR(tree) = 0.2$. Both the SDR and TMD agree with visual inspection: both indicate that m_0 is the best cluster, while m_2 is the worst and m_1 falls between them. Overall, both of the more sophisticated methods evaluate the tree far less charitably than the $RSPR$. This is perhaps to be expected, since m_2 is so sparsely strung throughout the tree, and since every module has at least 1 node deviating from the ideal of monomodularity.

Comparing Figure 24 to Figure 25-26, one can see that small differences between modularity and domain architecture tend to wash out in larger trees[11]. Figure 26 shows an almost negligible $RSPR(tree) = 0.081$, and the overall $SDR(tree)$ and $TMD(tree)$ both show improvements over those of Figure 24. Together, these results both confirm the visually apparent divide between the natural (m_1) and social sciences (m_0) present in this representation. This result adds weight to the claim that whether one i) considers merely the terminological overlap of discourses or ii) considers the groups formed by more advanced networks of inter-discourse terminological connection, one obtains a very similar result[12]. (See also Figures 28-29,

[10] The same result obtains with ‘Vinci_2’ and ‘Vinci_3’.

[11] Figure 25 is indeed a better representation of the underlying data than Figure 26, although since $TMD(tree)$ is calculated with respect to trees, Figure 24 must be compared to Figure 26.

[12] Figures 25-26 were analysed using the FastGreedy modularity algorithm of iGraph, see Figure

discussed below)

Network Representation

While networks can be drawn from the same type of distance data used to generate trees and hierarchies, they can also be used to represent the presence-absence data (used to calculate distance) directly. Since these data structures are often quite large in comparison to a simplified distance matrix, some limiting of the data is required before representation. Hive plots are used over conventional force-directed graphs to more clearly display inter-cluster structure, multiple connections between nodes are represented as thickness of the edge, and often an edge *weight* cutoff is employed to reduce the number of edges shown. Both of these parameters are specified in each figure caption. While there are plenty of interesting features of the following networks, I will attempt to confine myself to comments on their usefulness qua representation and to comparisons between other representations.

Taking Figure 27 as an example, it represents the Canonical+ dataset (as in Figures 8-11, 14-19). Comparing to Figure 24, which shows modularity mapped onto a tree-structure, the inter-module connections displayed in the hive plot can help us understand why the mapping of modularity onto a tree does not meet the ideal of monomodularity. We can see the connections between Sociology/Psychology (m_0) and Anthropology/Archaeology (m_2), partially explaining why the former two appear at the base of a branch containing the latter two in Figure 24.

The hive plot also facilitates quick comparison of connections to our intuitions. While the connection of Philosophy to Archaeology, Anthropology, Physics, and Cosmology seem obvious given the status of the former two as humanities and the physics-centrism of the philosophy of science, the connection to Meteorology strikes one as odd. In fact, this connection can be traced to the fact that meteorological writing began both in the Upanishads and with Aristotle, but only a cursory look at Figure 27 reveals that at least some (perhaps unintuitive) explanation of connection is required.

34 for a representation employing optimal modularity and subclustering. Both representations show a similar natural-science social-science divide, although Figure 34 does this in an albeit less binary way.

Hierarchical clustering of large datasets can also be usefully displayed on a hive plot. For comparison, consider Figures 28 and 29, which display the clustering and subclustering of the large FEGC dataset mapped onto a phylogenetic tree[13]. With a $TMD(tree^{fig28}) = 0.547$ and $TMD(tree^{fig29}) = 0.598$ it is apparent that here the subclustering is a worse fit to the tree. Nonetheless, one subcluster does achieve the ideal of monomodularity ($TMD(m_5^{fig29}) = 0.0$) not present in its supercluster. The following is an overview of the subclustering results with clusters (Figure 28) on the right and their subclusters (Figure 29) and intuitive names on the left (for $TMD(m_i)$ see figure legend).

$$m_0 \Longrightarrow \begin{cases} m_0 & \text{Genomics / Proteomics} \\ m_1 & \text{Biology Subdisciplines} \end{cases} \tag{6.1}$$

$$m_1 \Longrightarrow \begin{cases} m_2 & \text{Evolution} \\ m_3 & \text{Paleontology} \end{cases} \tag{6.2}$$

$$m_2 \Longrightarrow \begin{cases} m_4 & \text{Outliers} \\ m_5 & \text{Ecology} \end{cases} \tag{6.3}$$

While the above schema in conjunction with Figures 28-29 are simple enough, a summary of the clustering results are much more clearly presented in a hive plot, which represents clustering and subclustering simultaneously. In Figure 30 clusters are plotted along three radial axes, and the subclusters are plotted as coloured segments of each axis—no supplementary description is necessary.

Keeping in mind how the FEGC dataset was obtained (§ 5.4), it is interesting that Ecology related pages were so prominent that they emerged as dominating one of the main clusters (Figure 30, left). And while the main Ecology cluster has captured the outliers (including Chemistry) the subclustering in that axis shows a clear split between the two. The top axis also shows an interesting split between the 'omics' disciplines and a large cluster of subdisciplines of biology, and the rightmost cluster shows clearly the disproportionate relative size of the subclusters of m_2^{fig28}.

Besides mere ease of representation for clusters, a hive plot informs in ways not available from a (reticulated) tree. Since $g(v_i)$ for each node is represented as the

[13] Albeit, not a very well scoring tree, but it suffices for the explanation at hand.

size of each node, the most central nodes ($g(v_i) \gg 0$) in each cluster can be identified with relative ease (as diagrammed below).

$$m_0 \xrightarrow{g(v_i)\gg 0} \begin{cases} \text{Molecular Biology} \\ \text{Chemical Biology} \\ \text{Metagenomics} \\ \text{Systems Biology} \\ \text{Genomics} \end{cases} \tag{6.4}$$

$$m_1 \xrightarrow{g(v_i)\gg 0} \begin{cases} \text{Mathematical and Theoretical Biology} \\ \text{Structural Biology} \\ \text{Cell Biology} \\ \text{Computational Biology} \\ \text{Microbiology} \end{cases} \tag{6.5}$$

$$m_2 \xrightarrow{g(v_i)\gg 0} \begin{cases} \text{Evolutionary Developmental Biology} \\ \text{Evolutionary Biology} \\ \text{Evolutionary Psychology} \end{cases} \tag{6.6}$$

$$m_3 \xrightarrow{g(v_i)\gg 0} \left\{ \text{Paleontology} \right. \tag{6.7}$$

$$m_4 \xrightarrow{g(v_i)\gg 0} \begin{cases} \text{Zoology} \\ \text{Astrobiology} \end{cases} \tag{6.8}$$

$$m_5 \xrightarrow{g(v_i)\gg 0} \begin{cases} \text{Ecology} \\ \text{*Glossary of Ecology} \end{cases} \tag{6.9}$$

Of course, each of the main clusters can be further decomposed into plots showing internal connections, i.e. intra-cluster presence-absence data. These decompositions are presented for each axes in Figures 31-33. While Figure 31, (m_0 and m_1) is quite densely connected, one can now see more clearly the connections that ensure, for instance, that $g(\text{'Systems\_Biology'}) \gg 0$. Figure 32 ($m_2$ and m_3) shows a general increase and equalization of $g(v_i)$ values for each node excluding, understandably, the more niche disciplines of 'History_of_Paleontology' and 'Biosocial_Criminology'. In Figure 33 (m_4 and m_5) there has been an increase in the $g(v_i)$ of both 'Marine_Biology' and 'Population_Ecology' as compared to their centrality within the supercluster, indicating a more significant importance in establishing within-discipline connections. Interestingly, most subclusters show the same nodes with high $g(v_i)$ as appear within the superclusters, demonstrating that disciplines that are important between disciplinary groups tend to be important within them as well (But see the case of Philosophy and Psychology presented in Figures 34-35 and discussed below).

Let us finally turn our attention to the network representation of our largest canonical dataset in Figure 34. Firstly, the nodes with $g(v_i) \gg 0$ were often those included in the manually chosen Canonical dataset, although Canonical++ does not contain everything in the Canonical dataset (See Table 6). This provides some post facto justification for the choice of the Canonical set to begin with, and indicates another set of interest for future study, namely, the subset of Canonical++ where $g(v_i) \gg 0$.

The most striking feature of this network is the predominance of "top-down" connections (connections between pages about large groups of disciplines and their subdisciplines) between the left cluster and the top and right clusters. This group of connections far outweighs the number of connections between clusters roughly corresponding to natural (right) and social-sciences (top), lending further support to a pre-theoretic "two-domains" view of disciplines (See also Figures 25-26).

Subclustering of the right cluster shows a division between natural and pure-sciences, separating, for instance, 'Physics' and 'Chemistry' from 'Computer_Science' and 'Mathematics'. The uppermost cluster divides roughly into applied, public sector, and pure humanities in descending order. Figure 35 shows the subclustered hive plot of this cluster. Of immediate note is that while 'Psychology' shows a higher centrality than 'Philosophy' in the superclusters of Figure 34, this relationship is reversed in the subclusters of Figure 35, indicating that while 'Psychology' is important for establishing connections with disciplines outside of the humanities, 'Philosophy' is important for establishing them within.

While there are a variety of other possible analyses of the types described above, these analyses were chosen for their relative simplicity, intuitiveness, and their exemplification of the representational capacities of the structures of interest. Indeed, we have come a long way from the expressive power of intuitive hierarchical representations of the sciences.

Chapter 7

Conclusion

After so many representations have been offered, it would perhaps be desired to put my faith in one in particular. It would be easy enough to simply choose the representation (from the set offered) with the best fit between representation and data (it is Figure 19), or the one with the most expressive power (it is Figure 34). But this desiderata is complicated by external issues of ontology and internal issues of quality that are themselves informed by the analyses offered here (*"What do we add to the list of disciplines?" / "What do we analyse as a proxy for Disciplines?" / "How good is this proxy?"*).

Ontology

I have done my best to avoid any pontification about *which* discourses are disciplines, and which are not. I have partially circumvented this ontological problem by choosing "canonical" sets to work with and directing my analysis towards those in isolation. Nonetheless, this tactic was meant to circumscribe the problem of the structure of science within such a narrow, admittedly general and arbitrary, view of disciplines that analysis was *possible*—it was not meant as a decisive selection of the actual domain of the problem.

I prefer a much more deflationary approach to the ontological problem: disciplines are the things people work within when they create discourses that cluster better with themselves than with other discourses. These may descend (like species) from a lineage of work, or they might not. They might be given common and familiar names ('Ecology', 'Computer Science'), or they might not ('Developmental Evolutionary Psychology', 'Biological Steganography'). But in any case, the "structure of disciplines" or "structure of science" will depend on a prior selection of disciplines in general (and science in particular) that is well beyond the scope of any representational technique here deployed.

It would, all ontological issues considered, perhaps be better to insist that there really isn't such as thing as *the* structure of disciplines, or science. Instead, what can be offered, are structural representations of particular collections of scientific discourse—representations that are notwithstanding changes in language, new developments, and the plethora of conceptual and terminological ambiguities. While this may be unsatisfying, it is a degree more realistic than any Comptian hierarchy with gerrymandered disciplinary boundaries and deceptively simple orderings. In such a deflated and sufficiently weakened context, I feel confident that the representational and evaluative procedures applied here have generated many good representations.

Quality

Given any tree of disciplines examined here, it is possible to construct a natural language yet artificialy designed discourse that is sister to any given discipline[1]. Not to mention the fact that completely random collections of terms will have *some* structural relationship between them. This is hardly the kind of discourse about which we might want to form structural representations, not the meat of structural theorizing. But such artificial cases are not entirely different from the ways in which encyclopedic discourses are constructed in the public sphere[2].

Of course, Wikipedia and like encyclopedia serve as sources of disciplinary discourse in the same way that *D. melanogaster* serves as a source of human gene homologs. Wikipedia is a model case, capturing some of the simple relationships that would be difficult if not impossible (or immoral[3]) to study in the more complicated discursive context. Analyses of journal articles and SEP do circumvent some of the limitations of Wikipedia, but we must recognize that choosing structural representations involve choosing data of a given and limited quality.

Statistical assessment, here as everywhere, does allow us to make more fine grained distinctions between the quality of datasets. Having more or larger informative fraction of keywords, being rhizomatic in terminological overlap and having a

[1] Taking the nearest discipline d' to d, a discourse for a discipline sister to d could be constructed by a) removing any sentence in the discourse of d' that contained a term not in d and contained no term in d, or b) by adding a sentence to d with terms not already contained in d numbering less than those differing in d'.

[2] Indeed, it would be an entirely different project to attempt to quantify the amount of copy-and-paste sentence- or paragraph-transfers have affected the structure of Wikipedia.

[3] It would be nice to have every publication, unfinished or rejected manuscript, digitized lab bench notebook and private email.

high betweenness centrality for major disciplines are certainly well developed statistical notions of data quality. Indeed, it makes little sense to be any more confident in a representation than in the quality of the data used in its construction.

Two Analogies: 1916 and 1977

“[I]f I have succeeded in assigning linguistics a place among the sciences, it is because I have related it to semiology.”
F. de Saussure (1916), *Cours de Linguistique Générale*

Regardless of the correctness of the above, the more general claim that fashioning a place for a discipline in the sciences depends on finding its relations to other sciences is unquestionable—and perhaps even follows metaphorically from the expression ‘*a place for*’. So what of disciplinology? Allowing me some liberalities with the demarcation of disciplines, I would say that the project undertaken here is one of many that could be grouped under the heading of disciplinology; structural disciplinology is one of disciplinology’s sub-disciplines[4]. And disciplinology itself seems a sort of philosophy of science or bibliometrics.

So, if I have succeeded in assigning disciplinology a place in philosophy, it is because I have related it the philosophy of science. And if I have succeeded in assigning disciplinology a place in the sciences, it is because I have related it to bibliometrics.

***

“Phylogenetic relationships cannot be reliably established in terms of noncomparable properties. A comparative approach that can measure the degree of difference in comparable structures is required…Thus, comparative analysis of molecular sequences has become a powerful approach to determine evolutionary relationships…To determine relationships covering the entire spectrum of extant living systems, one optimally needs a molecule of appropriately broad distribution”
Carl R. Woese and George E. Fox (1977), *Phylogenetic structure of the prokaryotic domain: The primary kingdoms*

[4] Other subdisciplines perhaps include more traditional bibliometrics, questions of inter-disciplinary exchange and theoretical reductionism, questions dealing with the concept use specific to disciplines or inter-disciplinary conceptual difference.

Removing ‘Phylogenetic’ from the above and leaving “relationships cannot be reliably established in terms of noncomparable properties”, we are left with a truism about (structural) relationships[5]. Furthermore, when the relationships in question, or the representation desired, is significantly variable one requires comparisons to come in *degrees of difference.* Thus, when qualitative analysis fails to provide an understanding of the relationships between disciplines, a more *molecular* approach is required to establish relations. Finally, when the desired comparisons span the entire spectrum of extant systems, one requires data of *appropriately broad distribution.*

All disciplines will have a variety of qualitative features. But the discourses that end up represented in encyclopedic repositories are indeed of *appropriately broad distribution*, their sequence constraints (what Woese and Fox needed) makes them sufficiently *“molecular.”* They can be compared, and these comparisons come in degrees.

Together I think these analogies highlight the key contributions of this project: I carved out a place for disciplinology on the bibliometric side of bioinformatics, and I have done so with methods that, I like to think, would have appealed to the pioneers of phylogenetic bioinformatics itself.

Intuition Revisited

Surely only the most post-modern amongst us will be inclined to say something like, “Disciplines are nothing but texts”. So, when we interpret structural representations, like those I have shown today, as representing disciplines, concepts, or researchers we have to take this with a thick grain of salt. Nonetheless, when compared with attempts to derive such structures from memory, our intuitions about relatedness, complexity or consensus, I think it is obvious that an empirical approach enjoys a different standard of justification—if for no other reason than that empirically derived structures can be assessed statistically and compared with new data as it arrives.

With all the intuition-bashing I have been doing, it should be noted how fortunate it is that some features of the representations presented did indeed agree with some

[5] As Beiko notes, relationships can indeed be established between two entities even if they share no comparable properties as long as they share properties through other entities: establishing relationships transitively. Nonetheless, the fact remains that these intermediate comparisons *do* depend on the comparability of properties.

of our intuitions—no quantity of statistical data would be able to convince us that quantum mechanics is more closely related to 18th century art than to chemistry, or that sociology and biochemistry are sisters. And when we see the natural sciences group together to the exclusion of the social sciences, I think this gives us confidence, if not reason, to believe that there is indeed something natural about this grouping. Moreover, it is only once we begin to trust the representation to some extent that we can start picking out oddities (since without some faith, everything seems odd). Counter-intuitive, odd results are what trigger a closer investigation to see if there is a genuinely interesting reason for such an oddity to emerge.

Indeed, in demoting intuitive structures I did not mean to imply that something just being empirical is sufficient to make it better. Recall that, in connection with the canonical set of disciplines, I have only presented one kind of data, analyzed using one kind of distance metric, and although it was represented in many different structures, only one kind of phylogenetic analysis was used for each. Different data, distance metrics, and phylogenetic approaches can give mildly to wildly different answers—in fact, my early work on this project was mostly spent toiling around in the sandbox of algorithms, tools and data sources to find ones that did gave intuitive results! Put more concisely, structural representations of knowledge are model-dependent and interest-relative, and while this is what makes them useful, it is also what makes them difficult to interpret. Nonetheless, an empirical approach does enjoy certain "virtues of discovery", namely, that without an empirical approach one is not likely to find oddities[6] and irregularities that are unexpected given ones current stock of intuitions.

I hope to have convinced you that we are in a far more complex situation than one might be led to believe by intuitive hierarchical, disunity, or disorderly pictures—pictures of knowledge in general and science in particular. Nonetheless I believe there has been more disagreement than is necessary, since equivocation between, or dismissiveness of, different representational goals is often the cause. When our

[6] For example, early work on a larger collection of disciplines and concepts showed a connection between Demonology, dance and vegetarianism. Further investigation revealed that this was due to a mutual connection to Porphyry of Tyre, who wrote on all three topics and happens to be the locus of the first "Trees of Knowledge." The connection between Meteorology and Philosophy in Figure 27, produced by the Meteorological writings of Aristotle and in the Upanishads, would also qualify as such an oddity.

goals are clear, our methods appropriate, and means of evaluation and adjudication are offered, we can obtain good representations of the structure of knowledge that I believe help us understand the architecture of the complex conceptual world in which we live.

Appendix A

Figures

Figure A.1: Dataset: Canonical, Parameters: Unweighted, Rooted, Algorithm: UPGMA, $Evaluation : slope = 0.501$, $R^2 = 0.477$, $pvalue = 5.28e09$

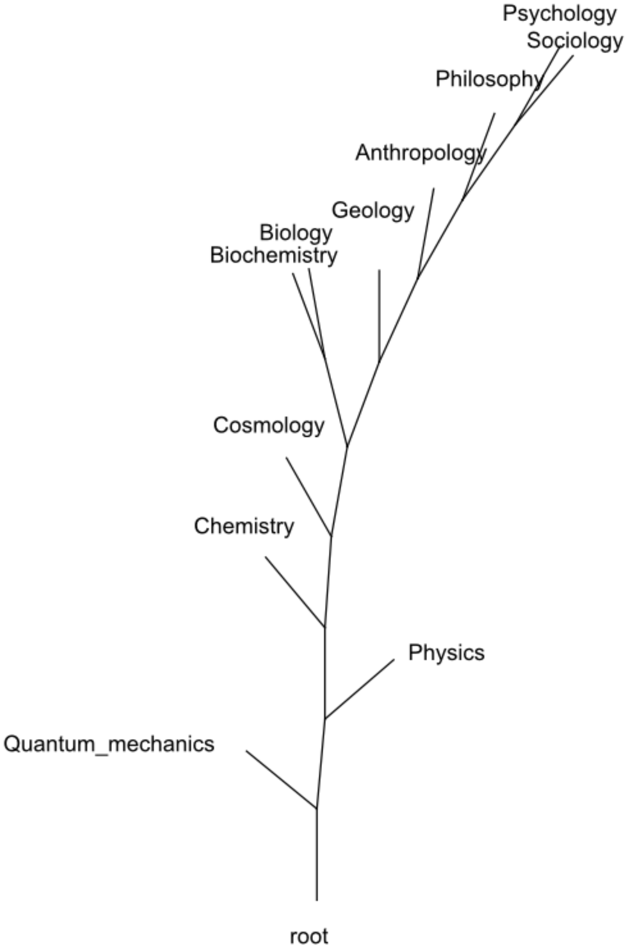


Figure A.2: Dataset: Canonical, Parameters: Weighted, Algorithm: UPGMA, $Evaluation : slope = 0.501$, $R^2 = 0.477$, $pvalue = 5.28e09$

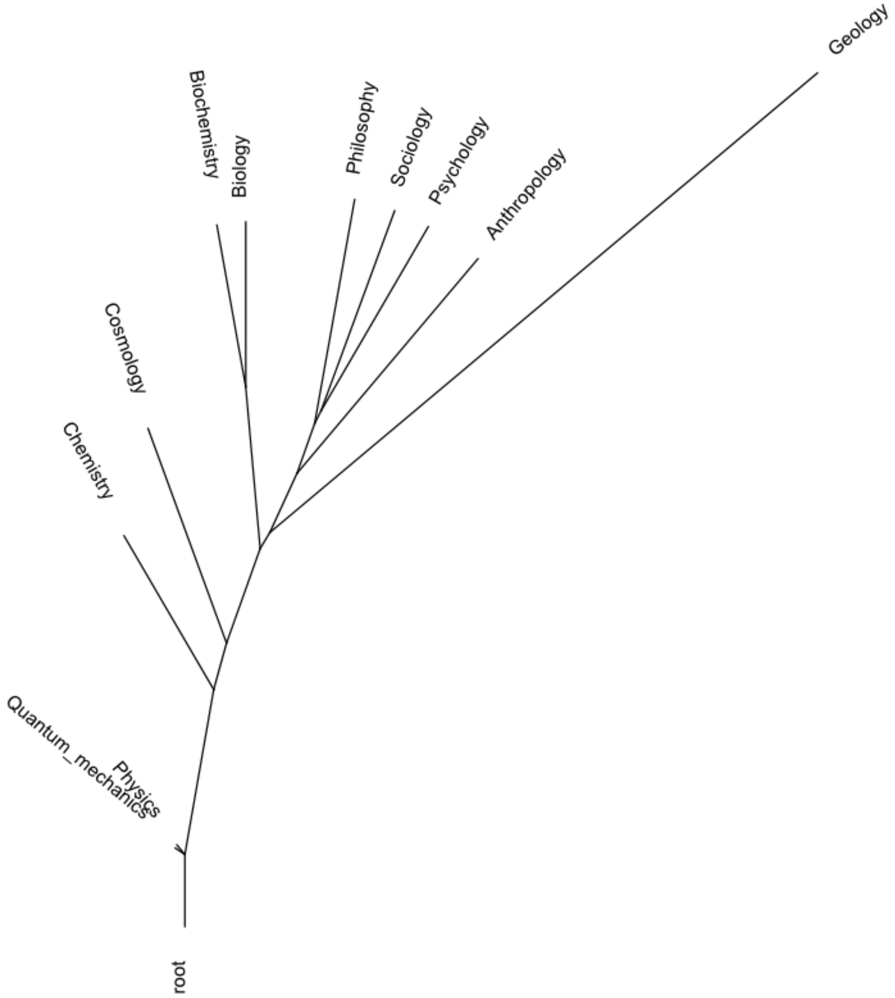


Figure A.3: Dataset: Canonical, Parameters: Unweighted, Algorithm: Neighbour-Joining (NJ), $Evaluation: slope = 0.969$, $R^2 = 0.712$, $pvalue = 5.82e16$

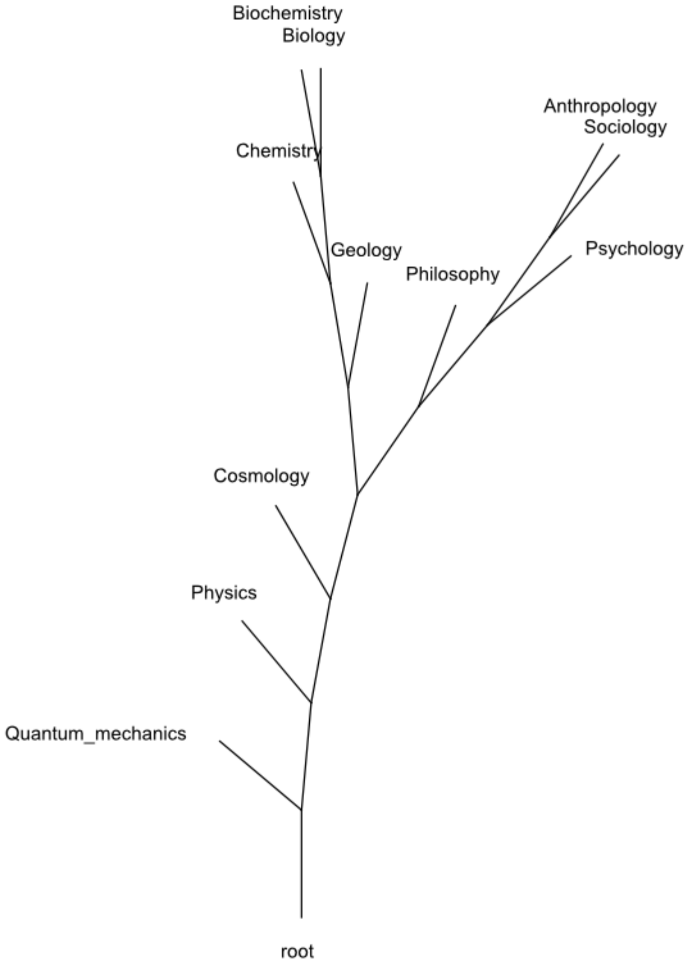


Figure A.4: Dataset: Canonical, Parameters: Unweighted, Unrooted, Algorithm: NJ, $Evaluation : slope = 0.969$, $R^2 = 0.712$, $pvalue = 5.82e16$

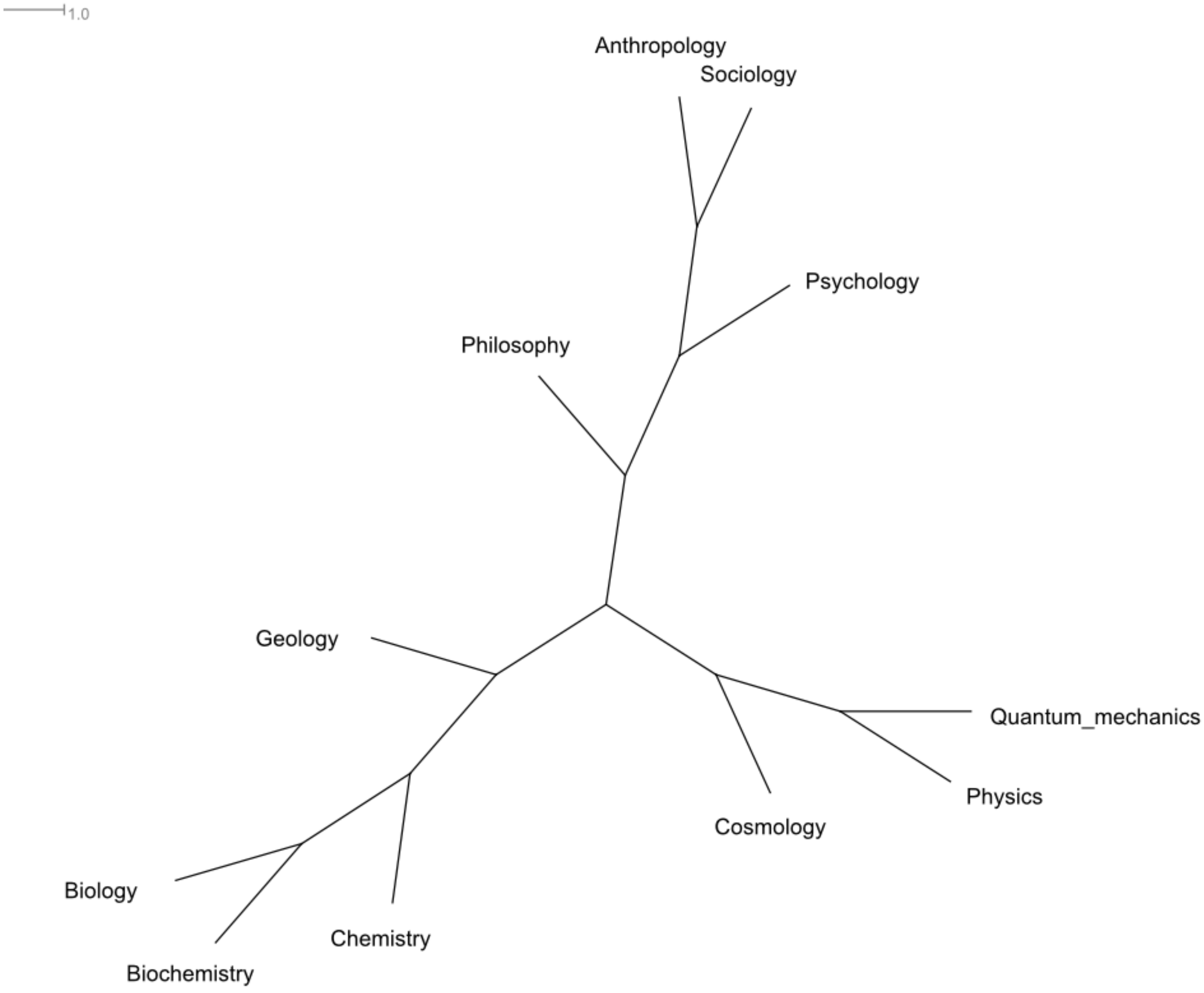


Figure A.5: Dataset: Canonical, Parameters: Unweighted, Manually Rooted in Philosophy, Algorithm: NJ, $Evaluation : slope = 0.969$, $R^2 = 0.712$, $pvalue = 5.82e16$

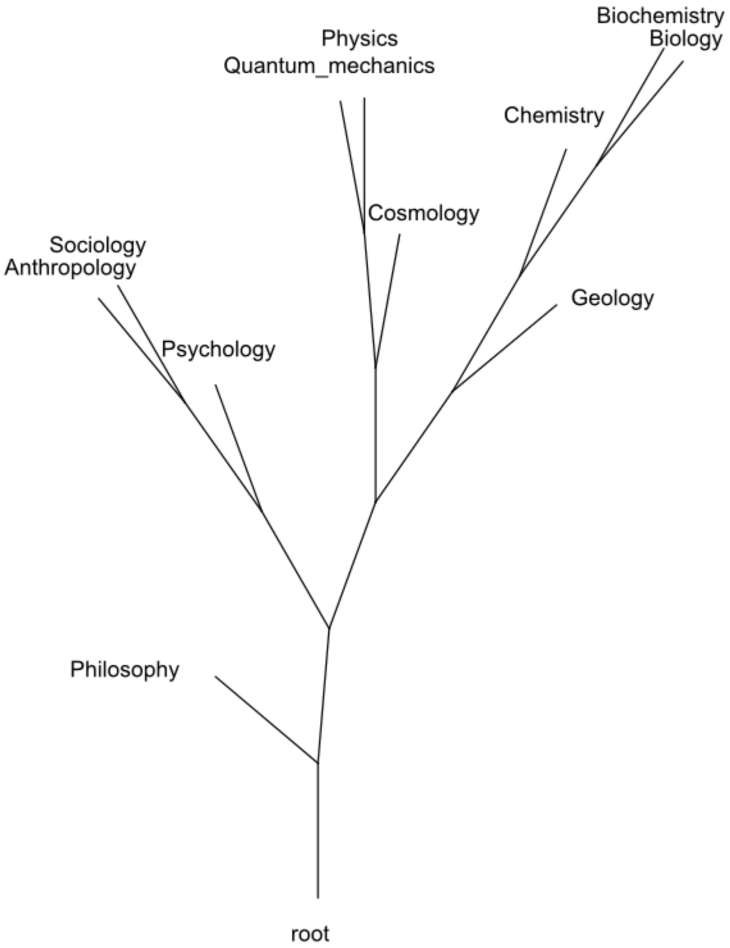


Figure A.6: Dataset: FEGC, Parameters: Slanted Cladogram, Algorithm: UPGMA, $Evaluation : slope = 0.729$, $R^2 = 0.388$, $pvalue = 0.0$

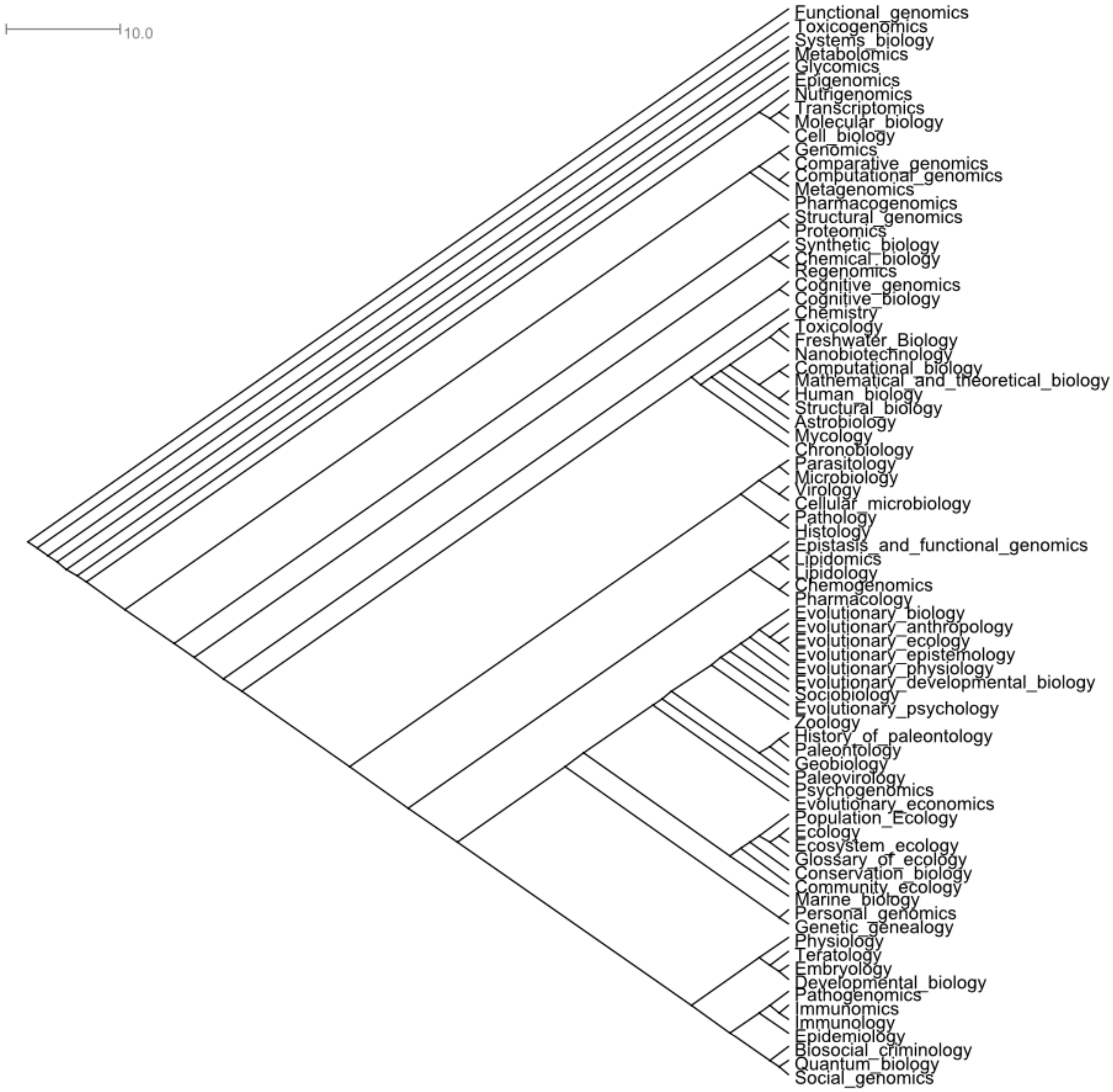


Figure A.7: Dataset: Canonical, Parameters: Phylogram, Algorithm: UPGMA, $Evaluation : slope = 0.501$, $R^2 = 0.477$, $pvalue = 5.28e09$

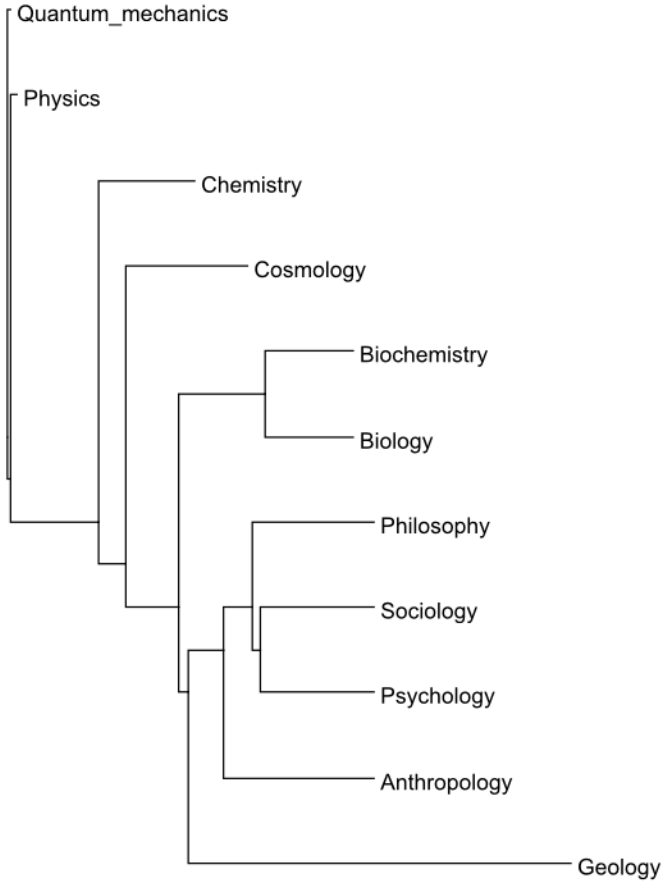


Figure A.8: Dataset: Canonical+, Parameters: Phylogram, Algorithm: UPGMA, $Evaluation : slope = 0.213$, $R^2 = 0.355$, $pvalue = 2.08e11$

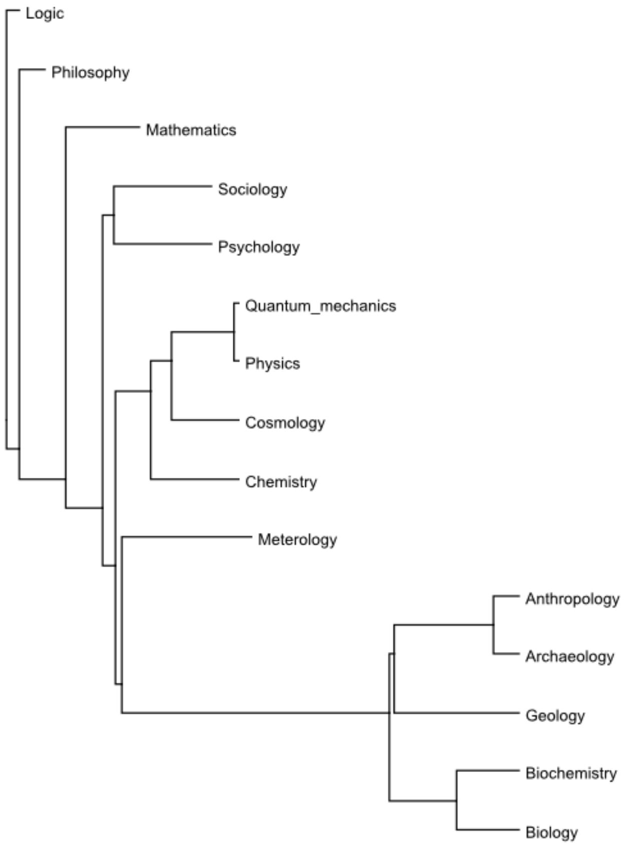


Figure A.9: Canonical+, Parameters: Phylogram, Algorithm: NJ, $Evaluation$: $slope = 1.005$, $R^2 = 0.771$, $pvalue = 8.64e35$

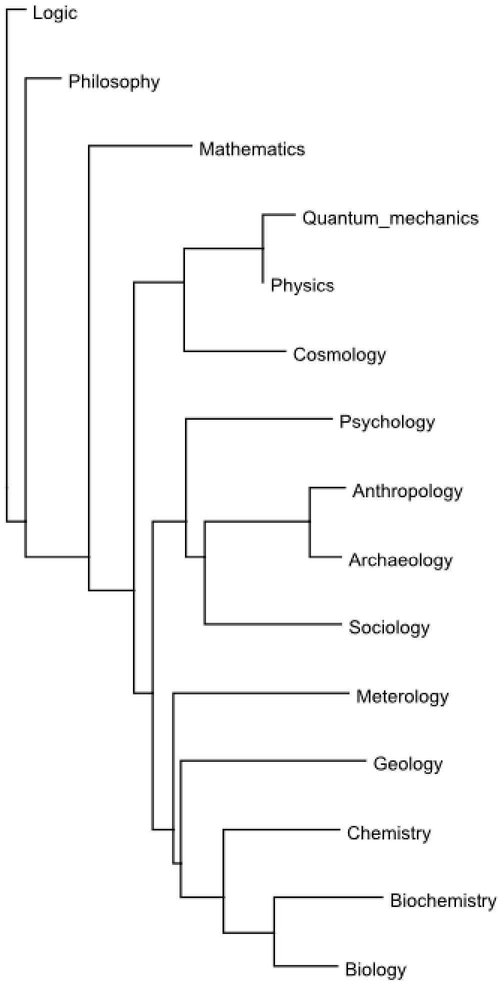


Figure A.10: Dataset: Canonical+, Parameters: Unweighted, Algorithm: NJ *Evaluation* : *slope* $= 1.005$, $R^2 = 0.771$, *pvalue* $= 8.64e35$ (Mantle test results: $R^2 = 0.771$, *pvalue* $= 0.001$)

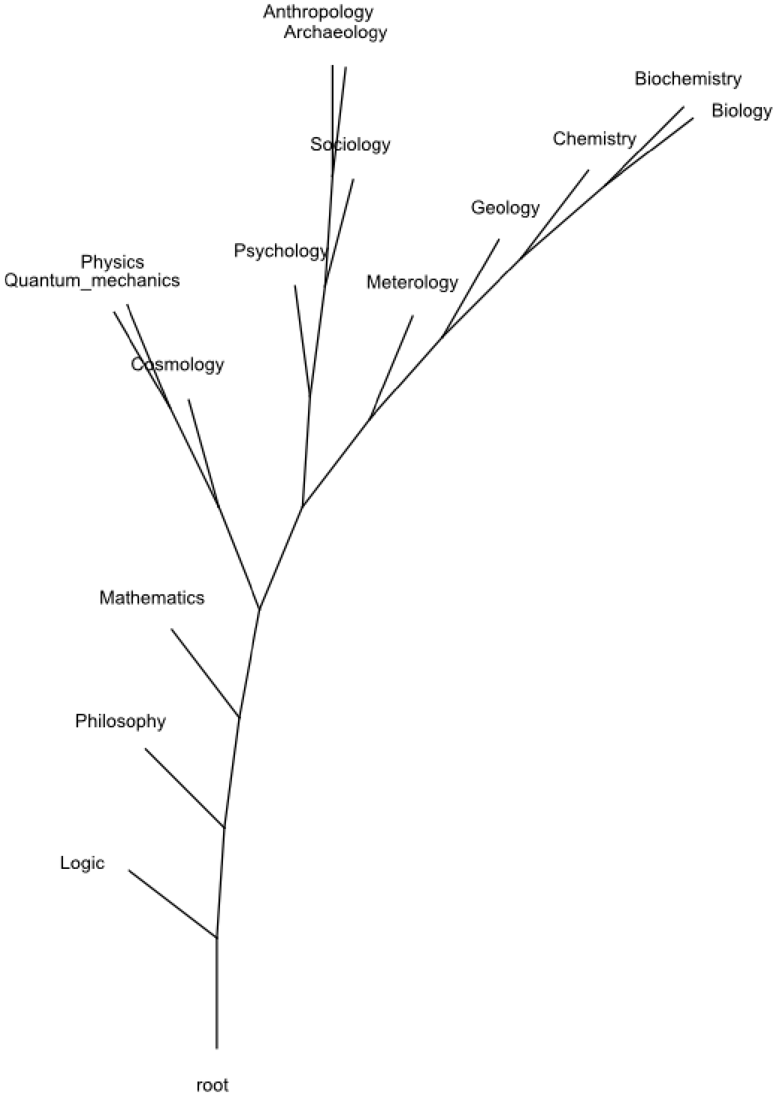


Figure A.11: Dataset: Canonical+, Parameters: Unweighted, Whole Page Parsing, Algorithm: NJ, $Evaluation: slope = 1.025$, $R^2 = 0.872$, $pvalue = 8.93e48$ (Mantle test results: $R^2 = 0.872$, $pvalue = 0.001$)

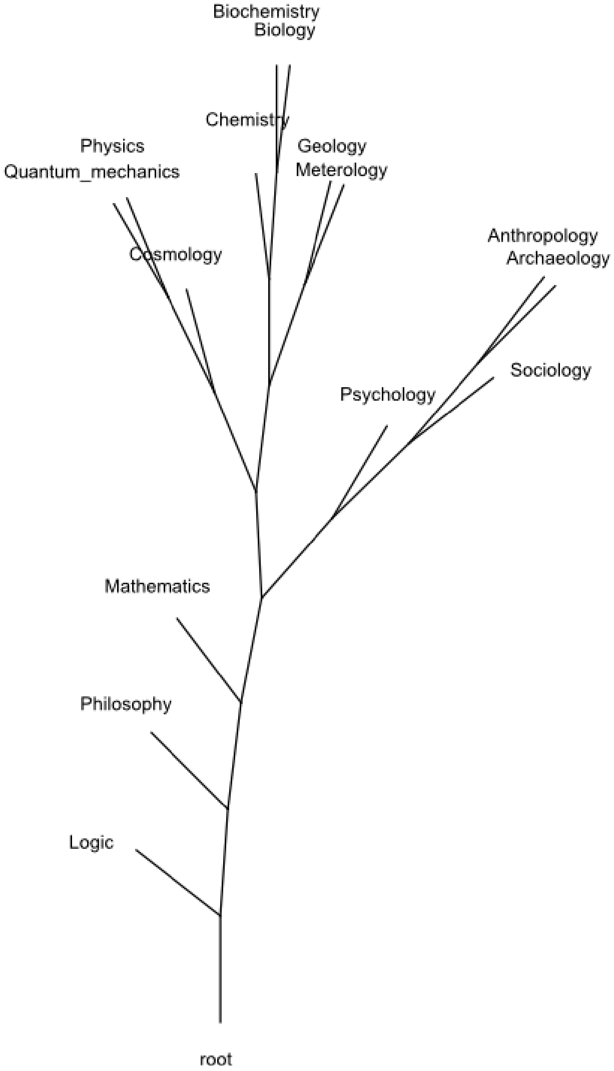


Figure A.12: Dataset: SEP-PhilBio, Parameters: Weighted, Algorithm: NJ, $Evaluation : slope = 1.03$, $R^2 = 0.892$, $pvalue = 1.21e32$

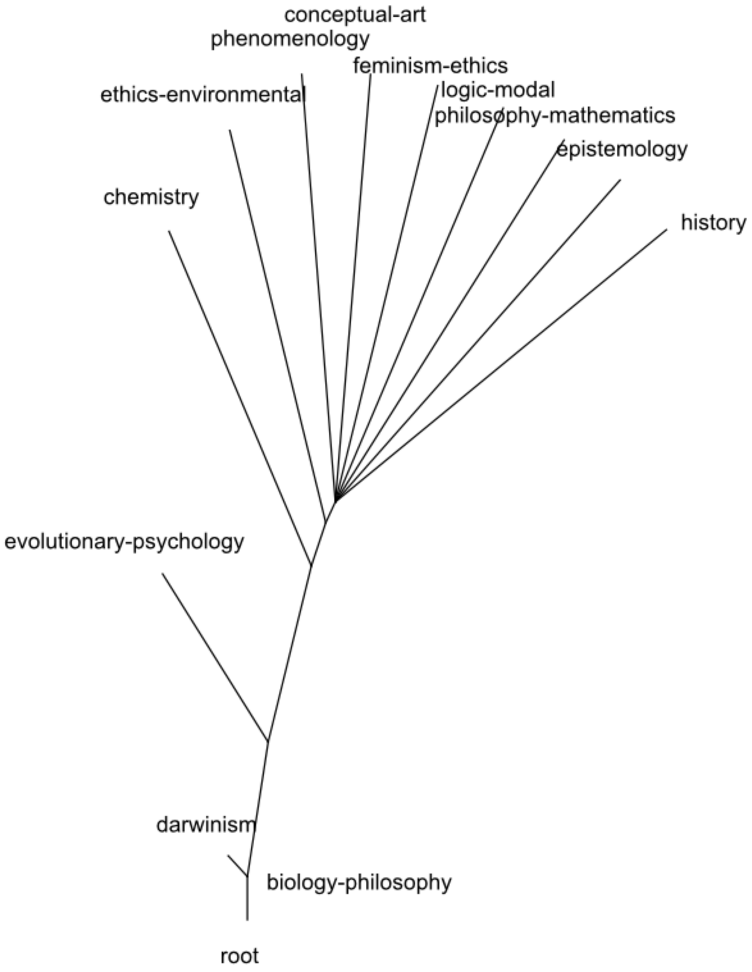


Figure A.13: Dataset: SEP-PhilBio, Parameters: Weighted, Whole Page Parsing, Algorithm: NJ, $Evaluation : slope = 0.974$, $R^2 = 0.785$, $pvalue = 4.68e23$

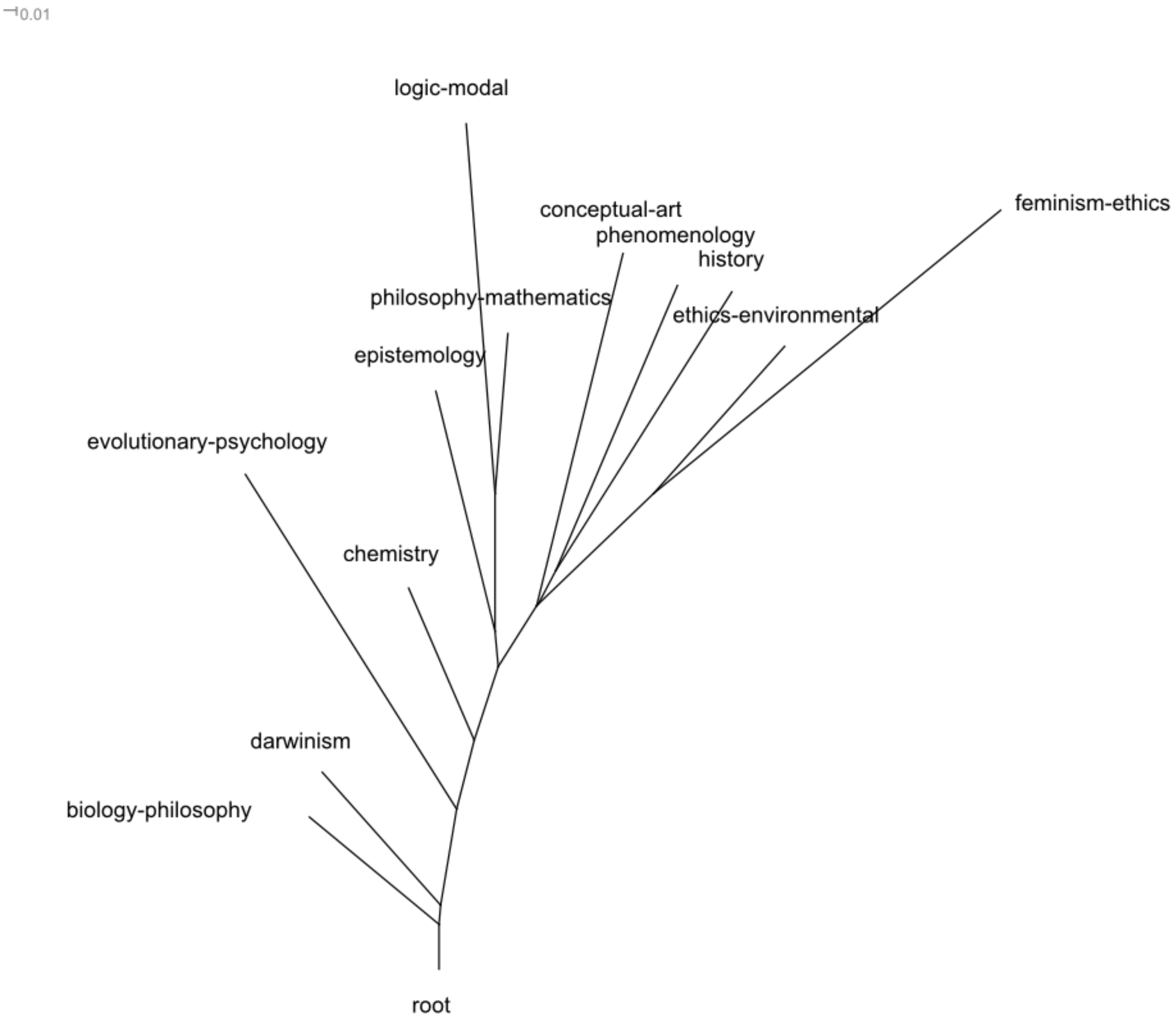


Figure A.14: Dataset: Canonical+, Parameters: Weighted, Rooted Algorithm: SplitsNetwork, $Evaluation : slope = 1.026$, $R^2 = 0.877$, $pvalue = 9.40e49$

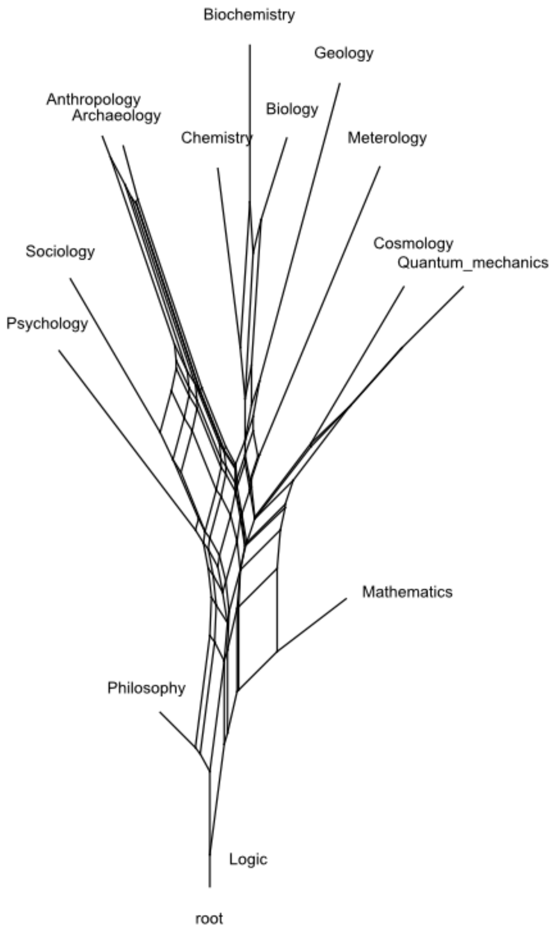


Figure A.15: Dataset: Canonical+, Parameters: Weighted, Algorithm: SplitsNetwork, $Evaluation: slope = 1.026$, $R^2 = 0.877$, $pvalue = 9.40e49$

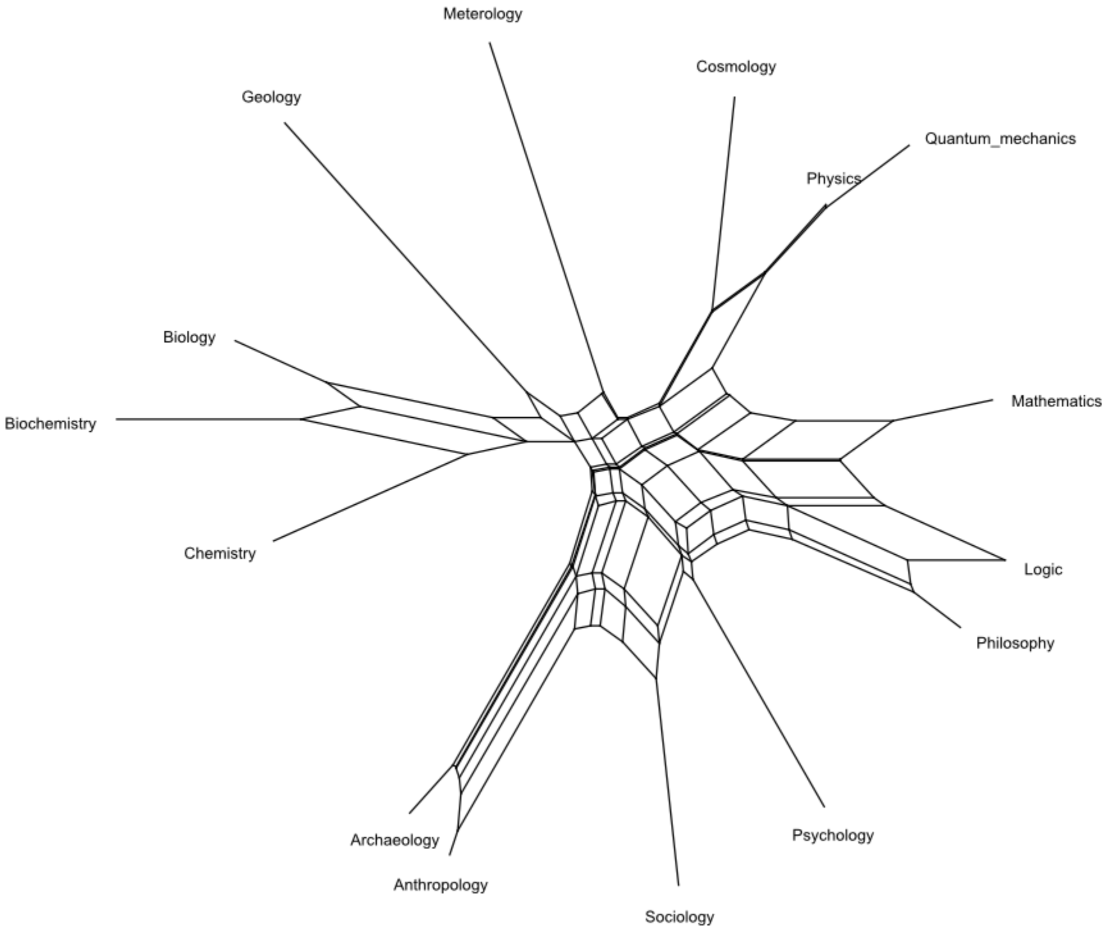


Figure A.16: Dataset: Canonical+, Parameters: Weighted, Algorithm: SplitDecomposition, $Evaluation : slope = 0.840$, $R^2 = 0.457$, $pvalue = 2.33e15$

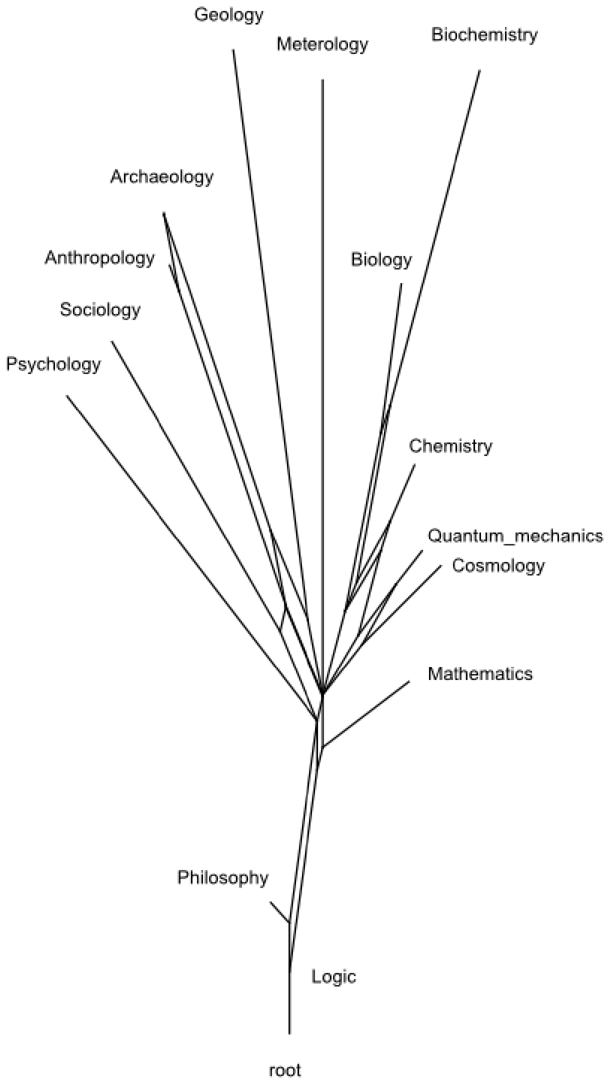


Figure A.17: Dataset: Canonical+, Parameters: Weighted, Whole Page Parsing, Algorithm: NJ, $Evaluation : slope = 1.024$, $R^2 = 0.872$, $pvalue = 8.93e48$

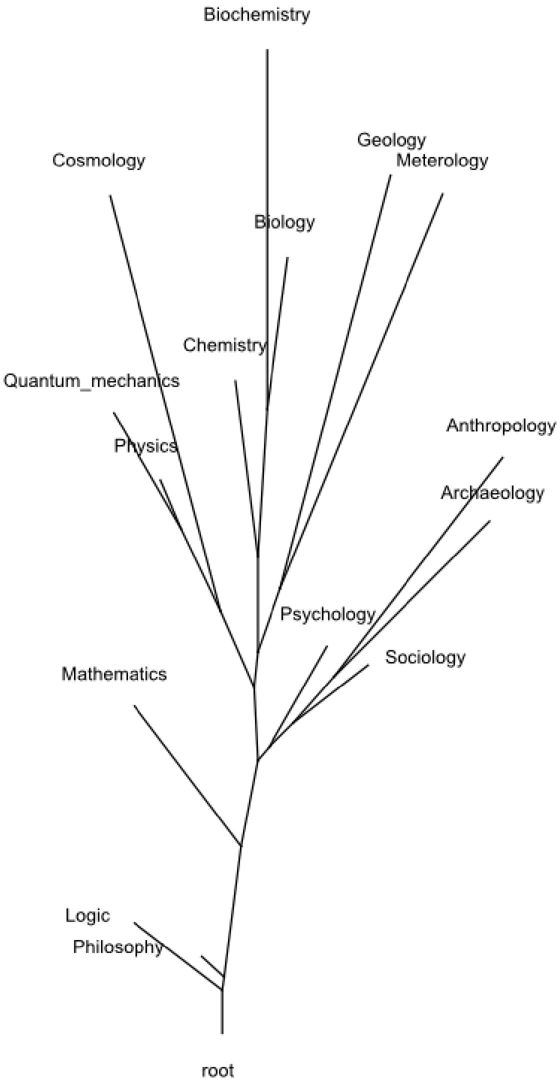


Figure A.18: Dataset: Canonical+, Parameters: Weighted, Whole Page Parsing, Rooted Algorithm: SplitsNetwork, $Evaluation : slope = 1.007$, $R^2 = 0.936$, $pvalue = 2.92e63$

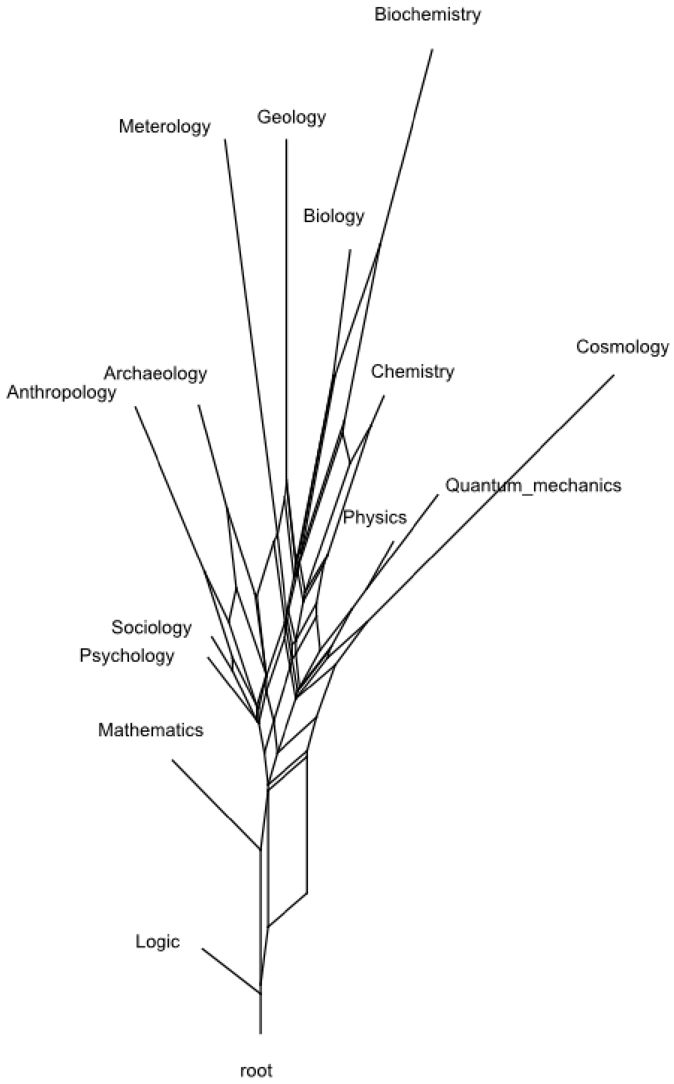


Figure A.19: Dataset: Canonical+, Parameters: Weighted, Whole Page Parsing, Algorithm: SplitsNetwork, $Evaluation$: $slope$ = 1.007, R^2 = 0.936, $pvalue$ = $2.92e63$

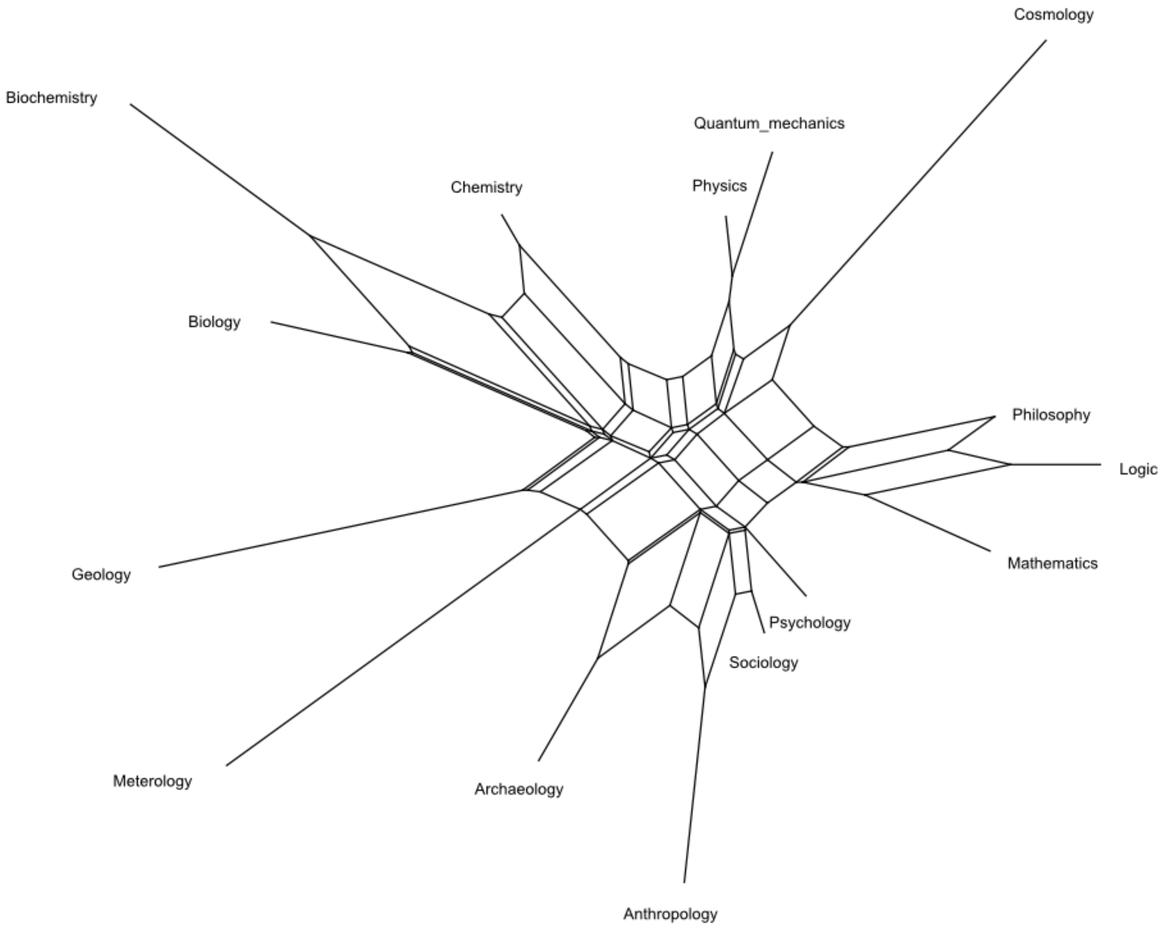


Figure A.20: Dataset: CGEB Papers, Parameters: Weighted, Whole Page Parsing, Algorithm: NJ, $Evaluation : slope = 1.03$, $R^2 = 0.778$, $pvalue = 0.001$

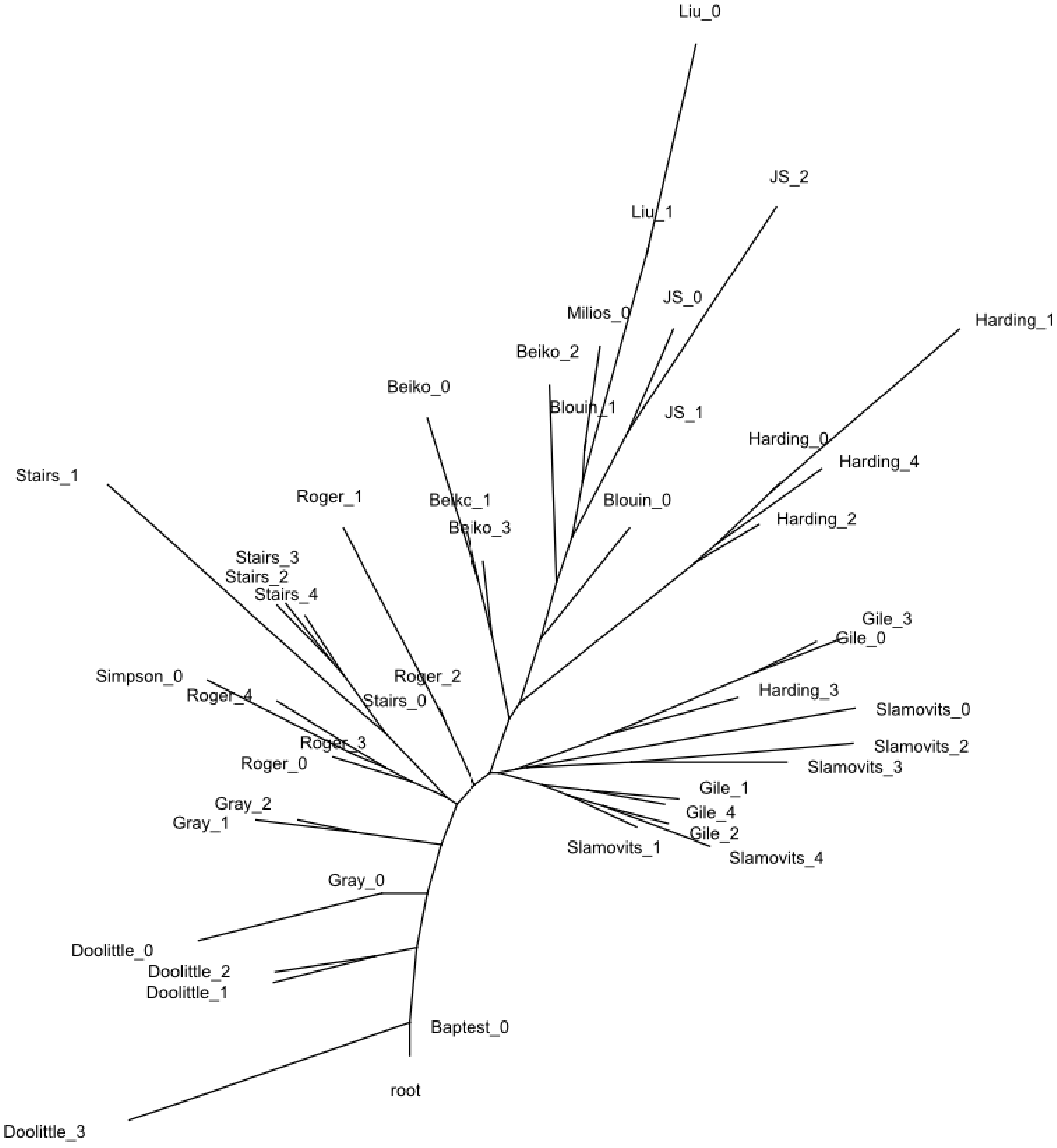


Figure A.21: Dataset: CGEB Papers, Parameters: Weighted, Whole Page Parsing, Algorithm: SplitsNetwork, $Evaluation : slope = 1.022$, $R^2 = 0.852$, $pavlue = 0.001$

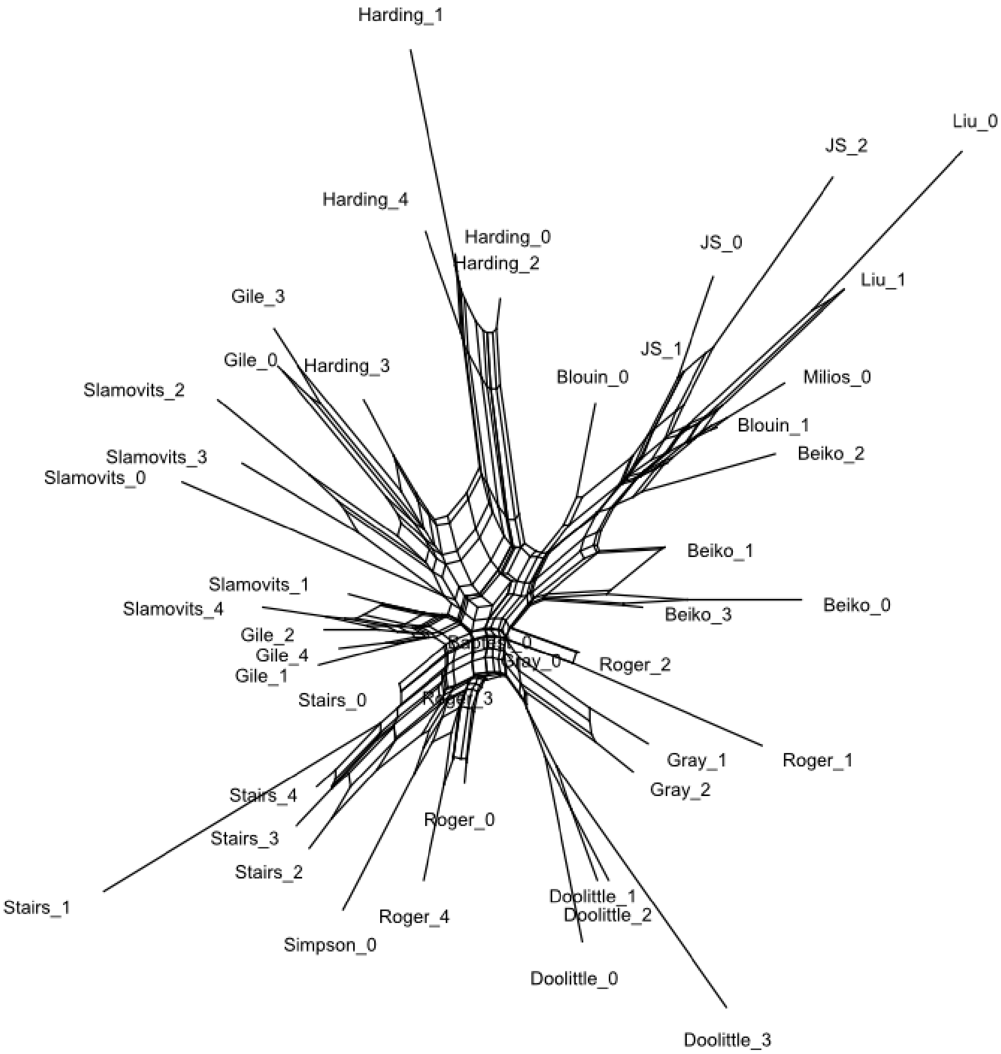


Figure A.22: Tree of Dalhousie Philosophy Department: $R^2 = 0.798$

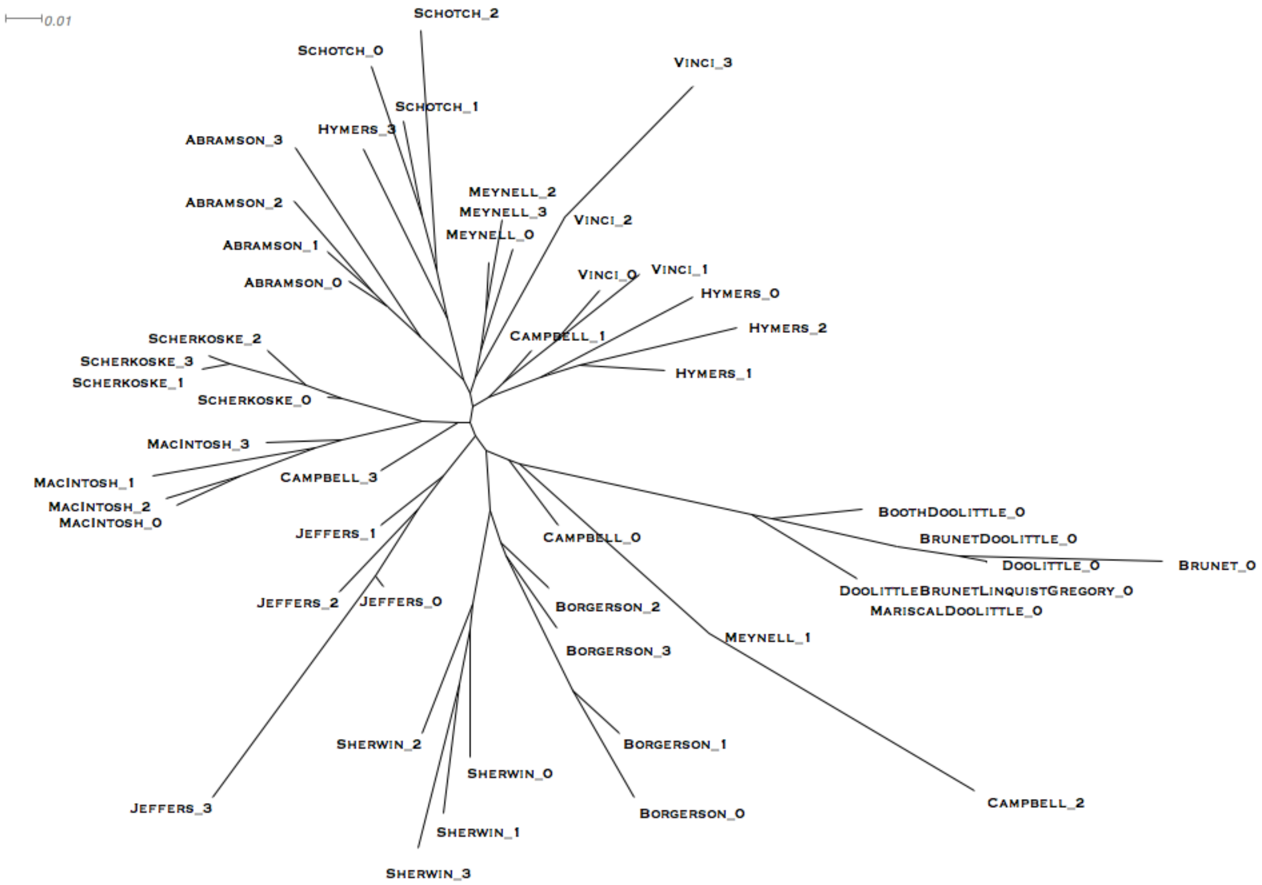


Figure A.23: Network of Dalhousie Philosophy Department: $R^2 = 0.83$

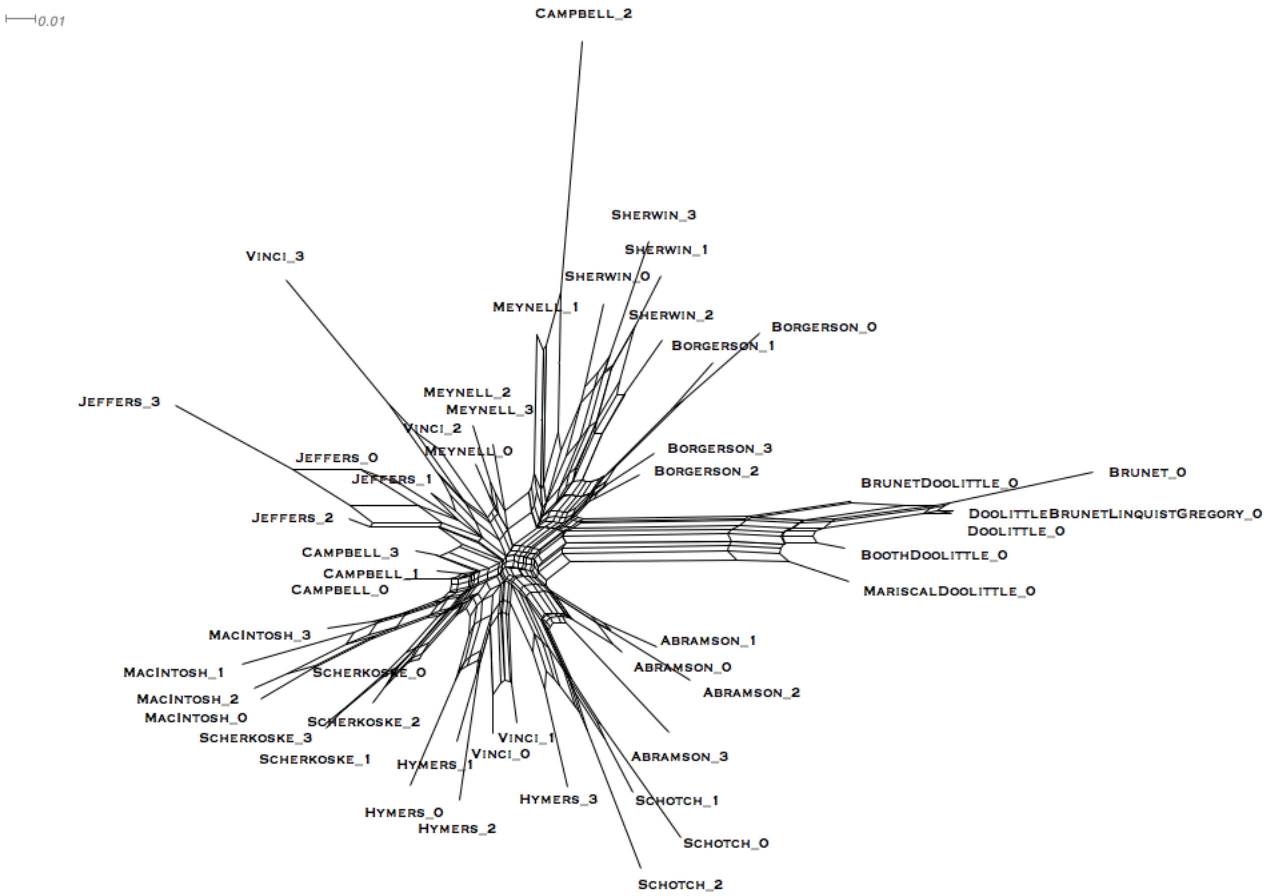


Figure A.24: Dataset: Canonical+, Parameters: Optimal Modularity, Weighted, Whole Page Parsing, Algorithm: NJ. $Evaluation$: $slope = 1.005$, $R^2 = 0.771$
$RSPR(tree) = 3/15 = 0.2$
SDR:
$SDR(m_0) = 0.533$
$SDR(m_1) = 0.901$
$SDR(m_2) = 1.128$
$SDR(tree) = 0.854$
TMD:
$TMD(m_0) = 0.187$
$TMD(m_1) = 0.533$
$TMD(m_2) = 0.605$
$TMD(tree) = 0.441$

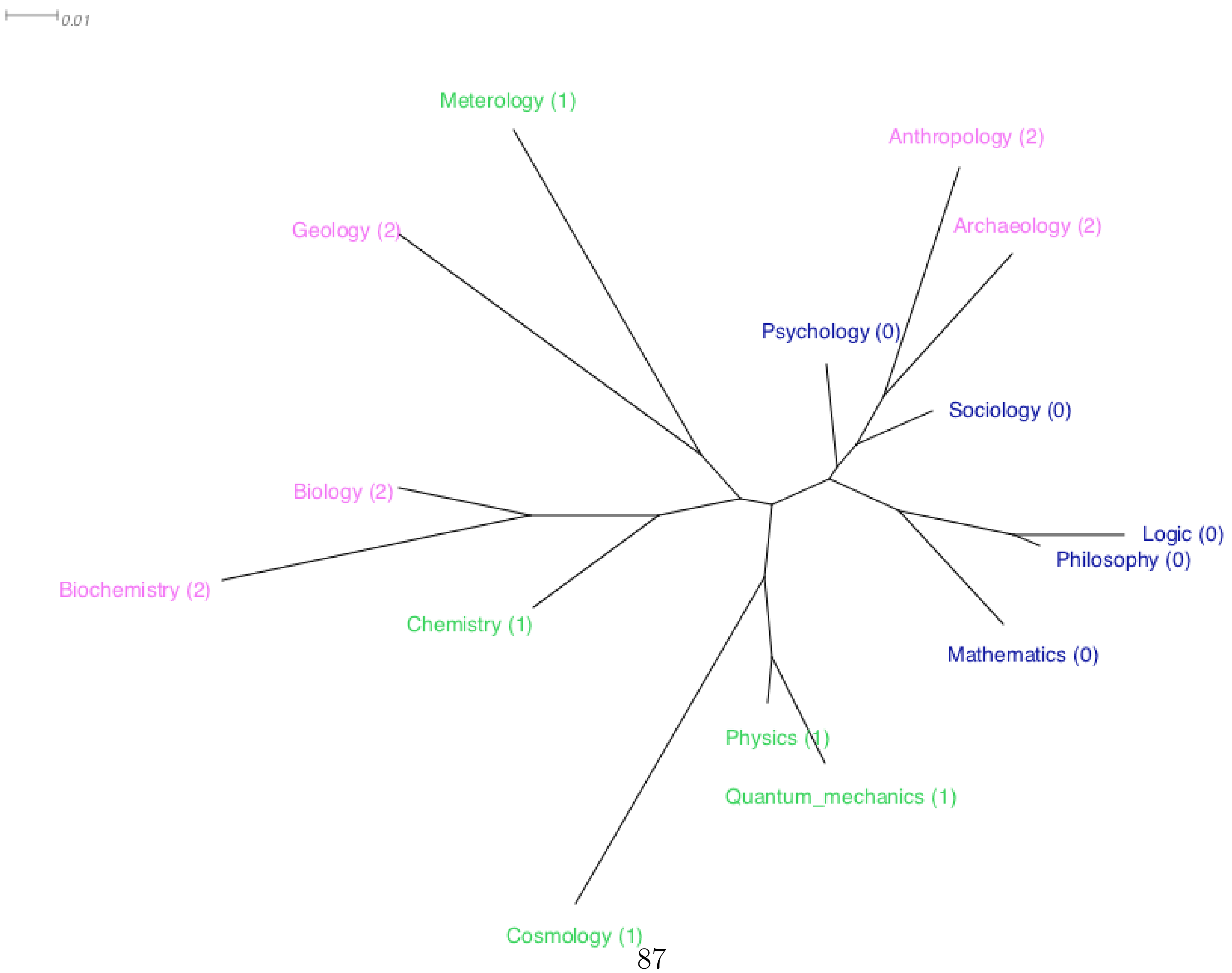


Figure A.25: Dataset: Canonical+, Parameters, Optimal Modularity, Weighted Whole Page Parsing, Algorithm: SplitsNetwork. Shows approximate natural Vs social science division and network representation is preferred. $Evaluation : slope = 1.028,\ R^2 = 0.76$

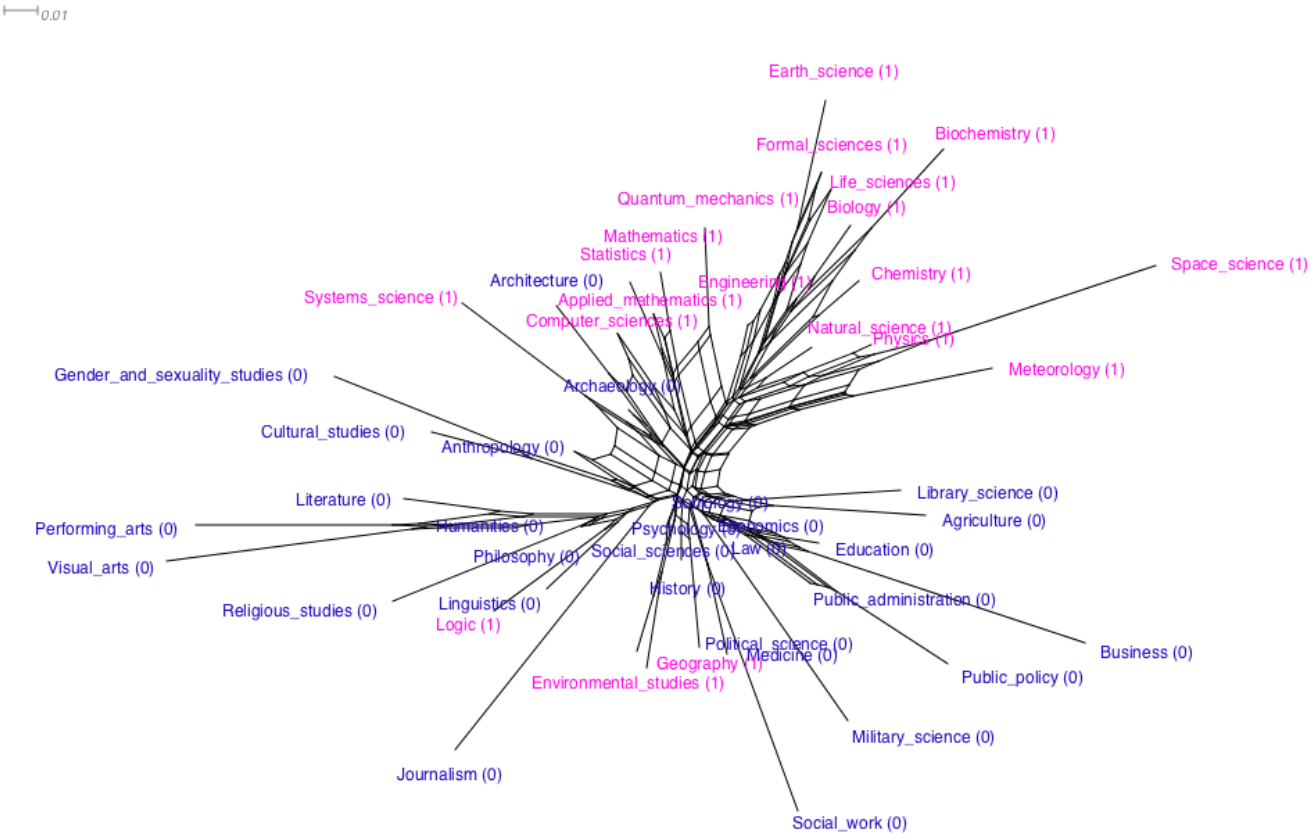


Figure A.26: Dataset: Canonical++, Parameters, Optimal Modularity, Weighted Whole Page Parsing, Algorithm: NJ. Shows approximate natural Vs social science division. Network representation is preferred. $Evaluation : slope = 1.117$, $R^2 = 0.718$

$RSPR(tree) = 4/49 = 0.081$

SDR:

$SDR(m_0) = 0.832$

$SDR(m_1) = 0.903$

$SDR(tree) = 0.868$

TMD:

$TMD(m_0) = 0.384$

$TMD(m_1) = 0.409$

$TMD(tree) = 0.396$

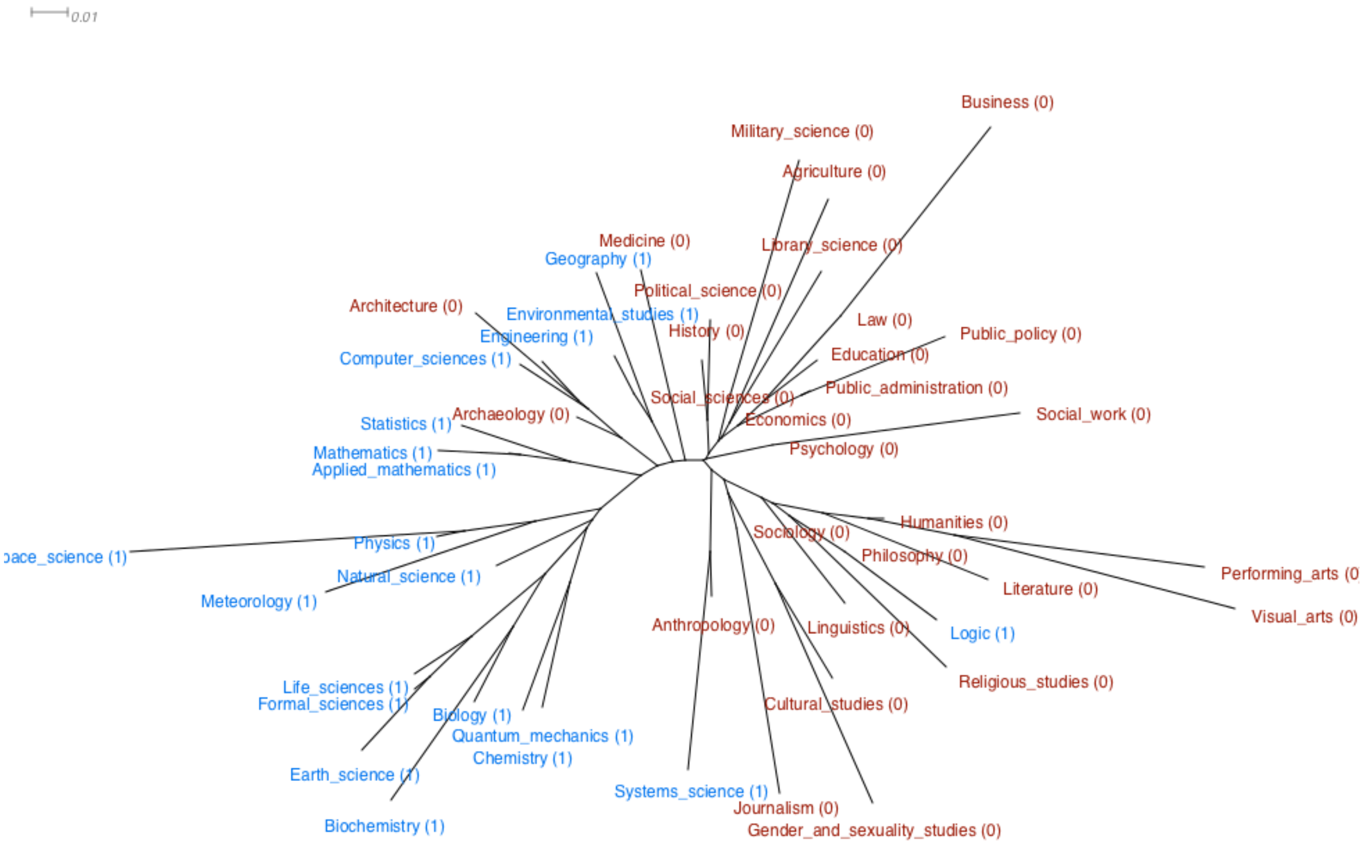


Figure A.27: Hive plot of Canonical+ dataset with modularity computed using iGraph optimal modularity and connections between nodes filtered to include only connections with *weight* ≥ 10, and $thickness = \sqrt[2]{weight}$ (default unless otherwise specified)

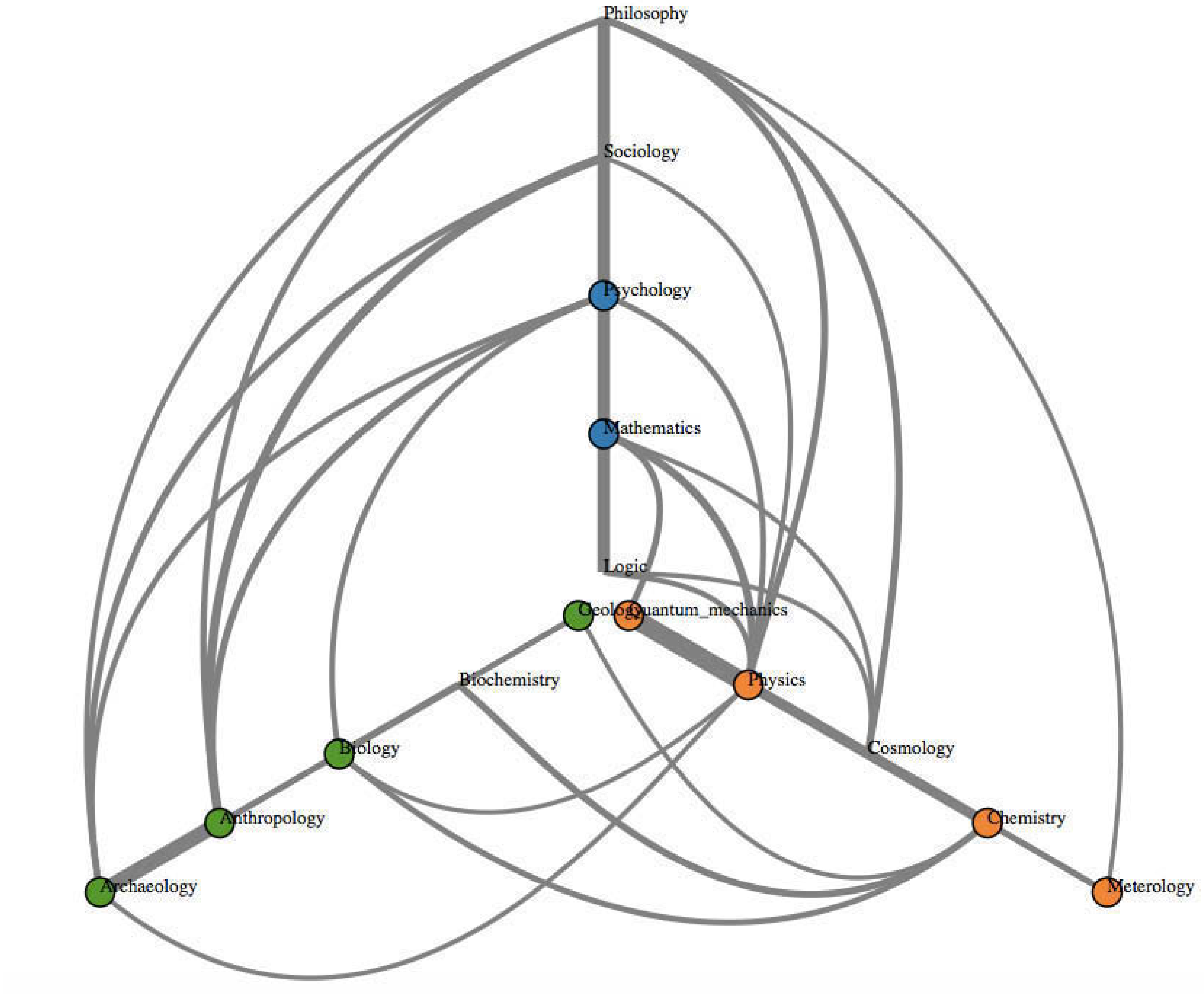


Figure A.28: Dataset: FEGC, Parameters: Slanted Cladogram, Algorithm: UPGMA, $Evaluation : slope = 0.729$, $R^2 = 0.388$, $pvalue = 0.0$
$TMD(m_0) = 0.226$
$TMD(m_1) = 0.745$
$TMD(m_2) = 0.672$
$TMD(tree) = 0.547$

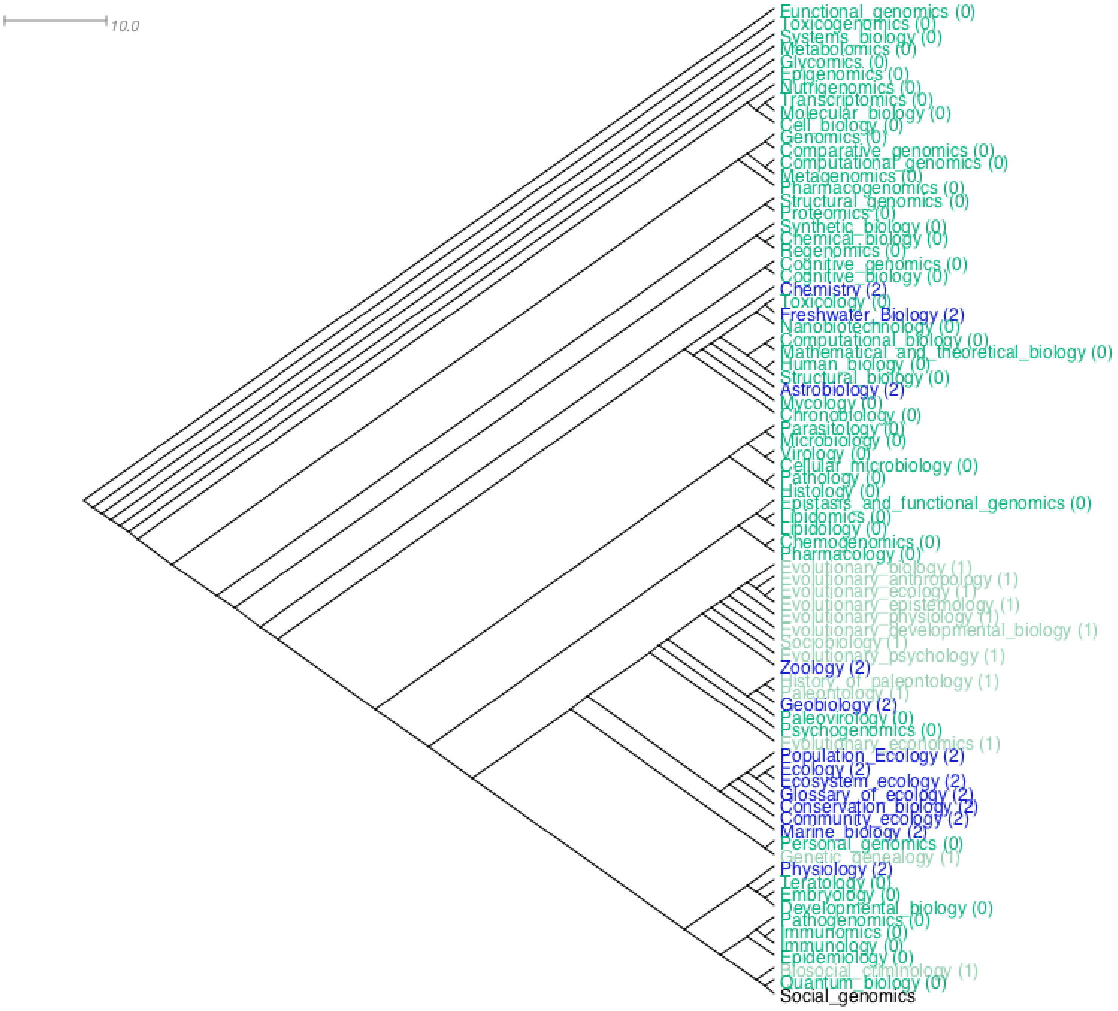


Figure A.29: Dataset: FEGC, Parameters: Slanted Cladogram, Algorithm: UPGMA, $Evaluation : slope = 0.729$, $R^2 = 0.388$, $pvalue = 0.0$
$TMD(m_0) = 0.548$
$TMD(m_1) = 0.513$
$TMD(m_2) = 0.708$
$TMD(m_3) = 0.816$
$TMD(m_4) = 1.000$
$TMD(m_5) = 0.000$
$TMD(tree) = 0.598$

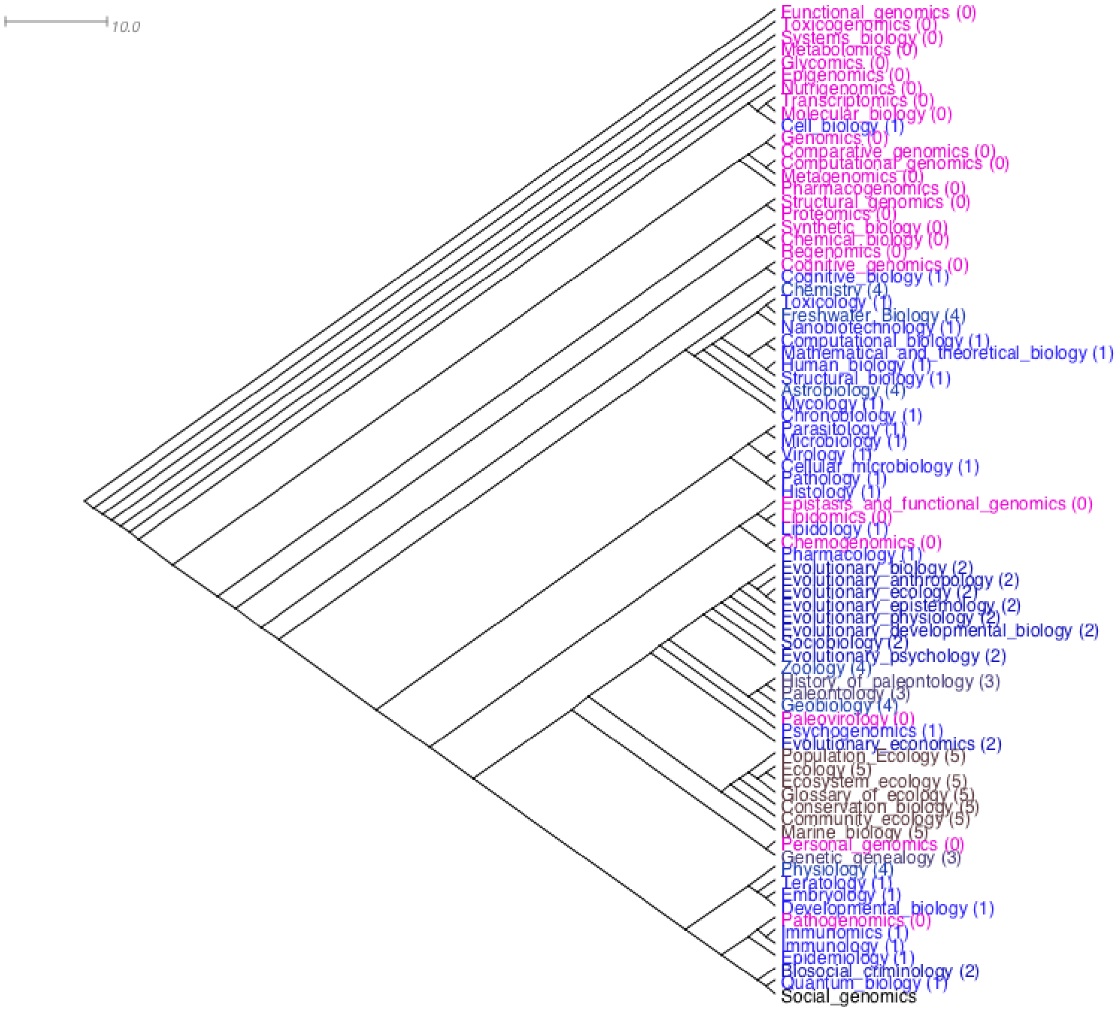


Figure A.30: Hive plot of FEGC dataset, optimal modularity and subclustering, default filtering.

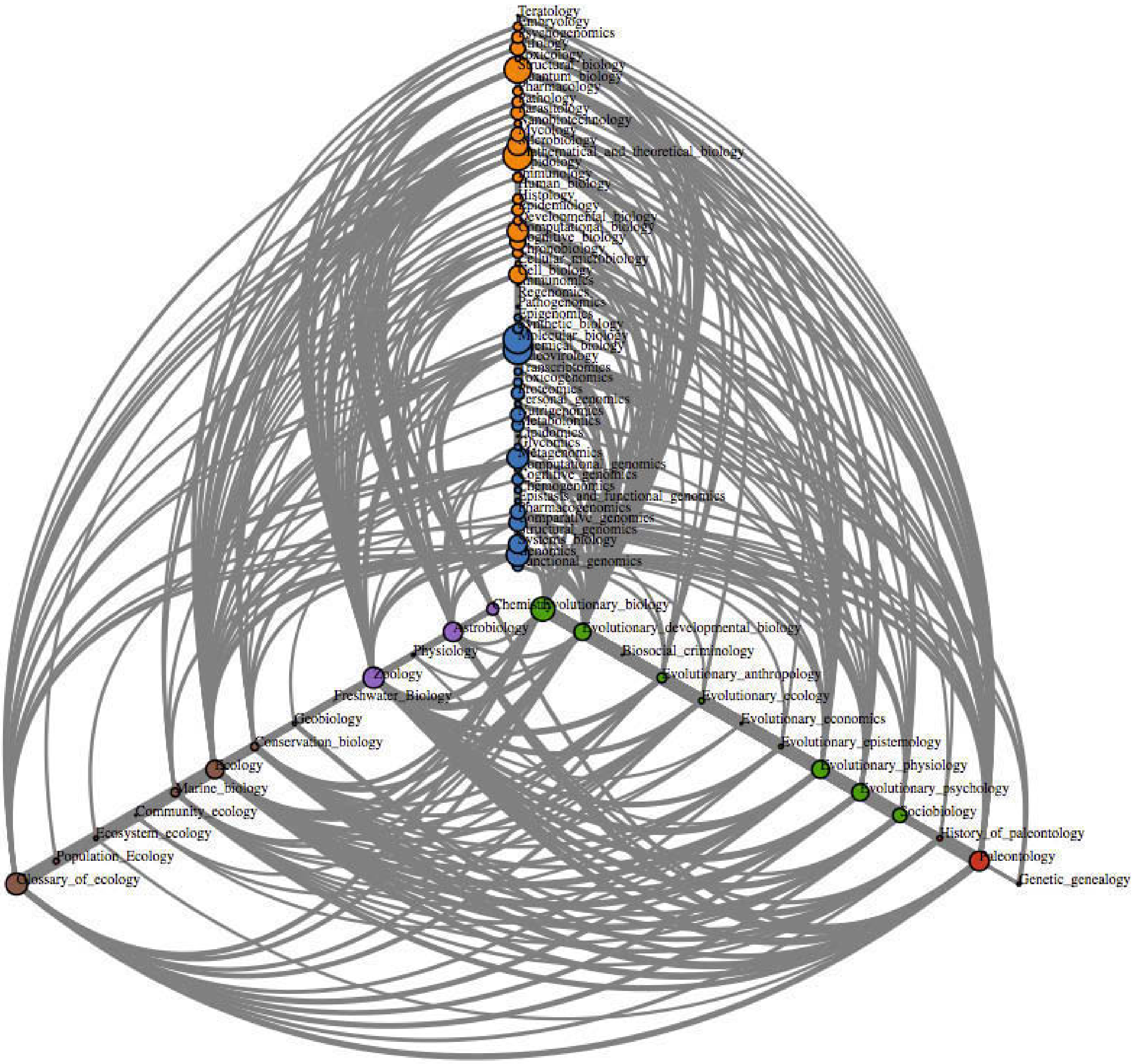


Figure A.31: Hive plot of subclustering of top cluster for Figure 30

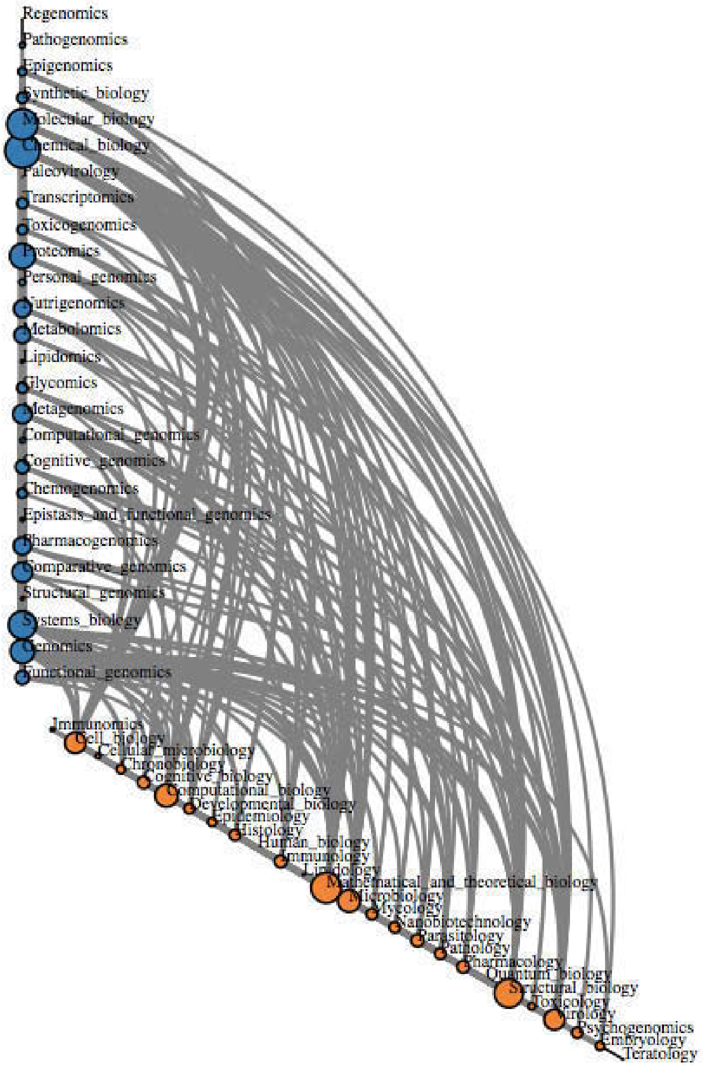


Figure A.32: Hive plot of subclustering of right cluster for Figure 30

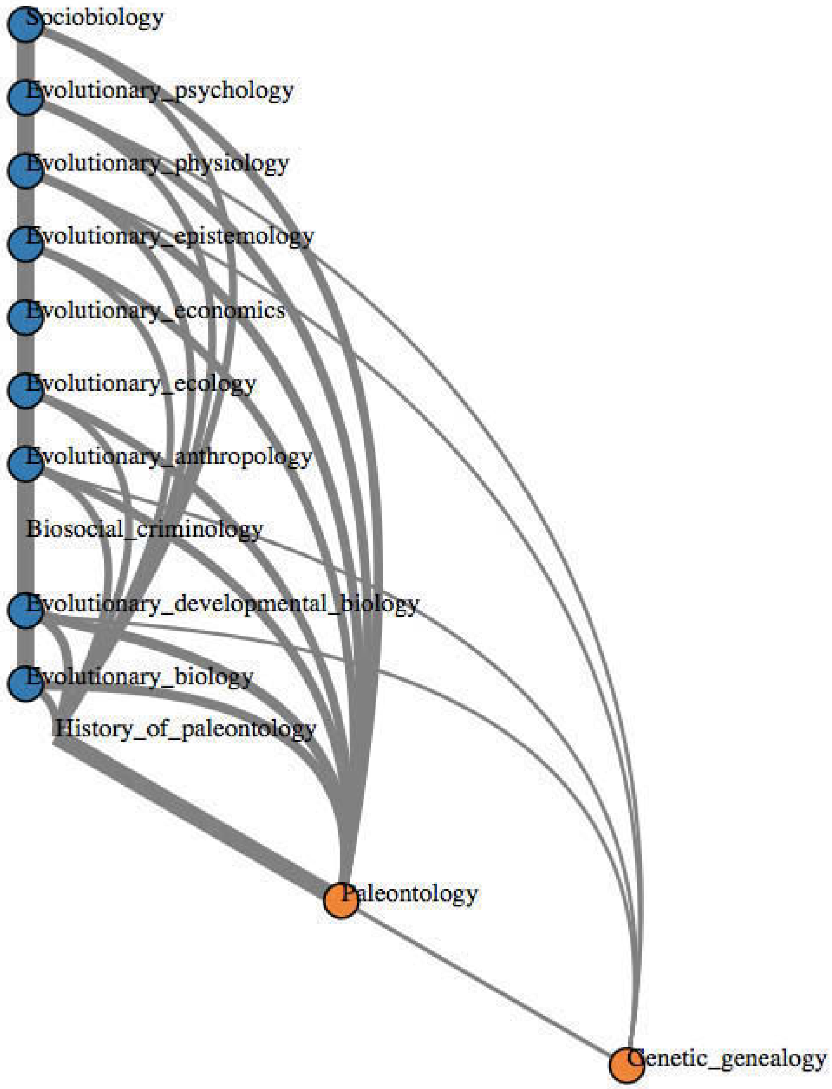


Figure A.33: Hive plot of subclustering of left cluster for Figure 30

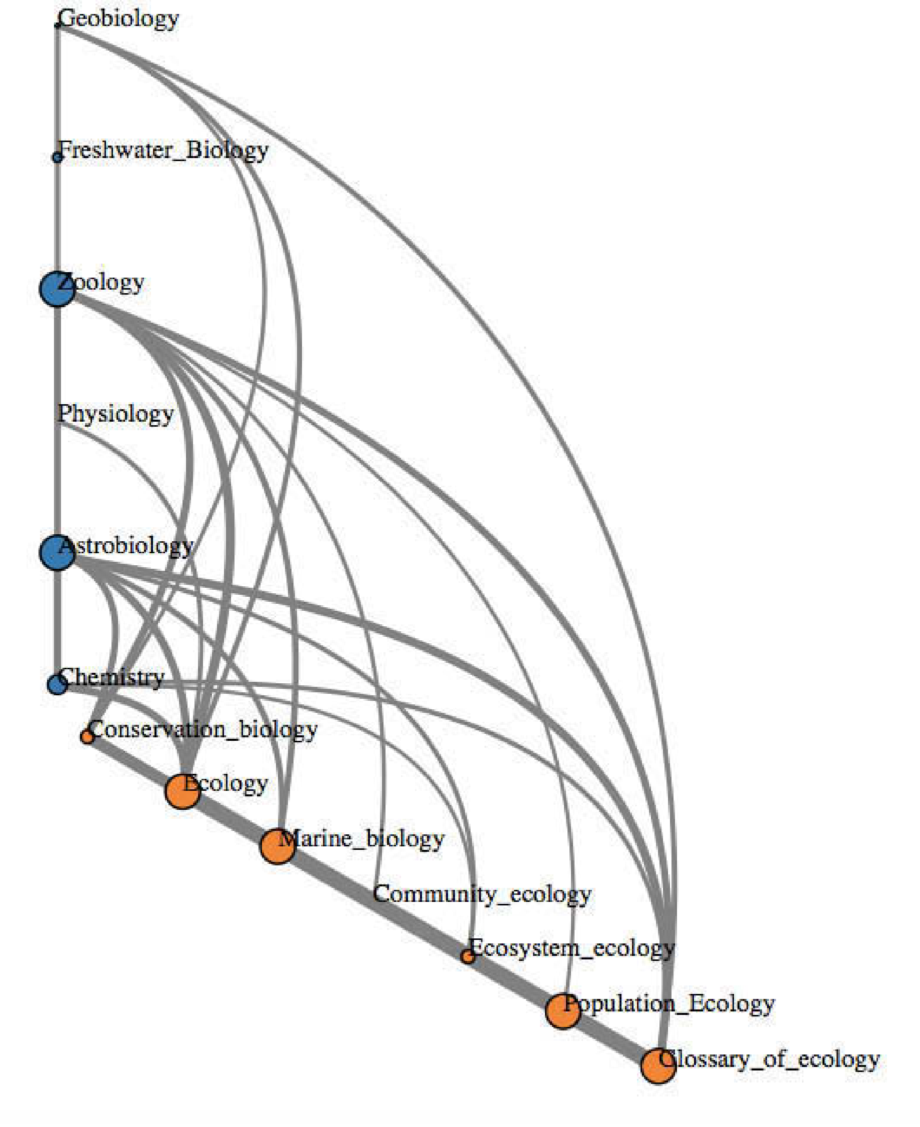


Figure A.34: Hive plot of subclustering of Canonical++ dataset

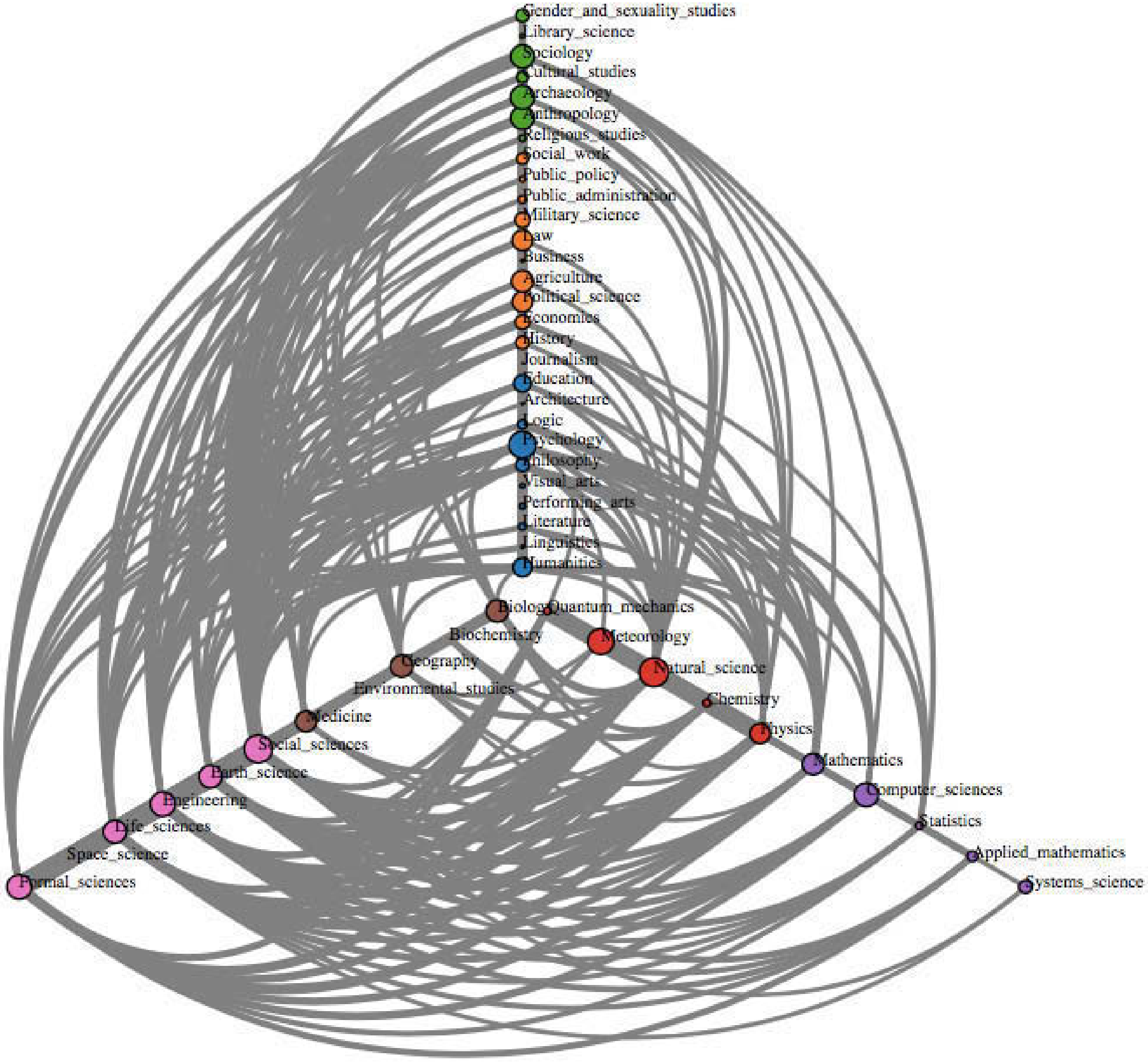
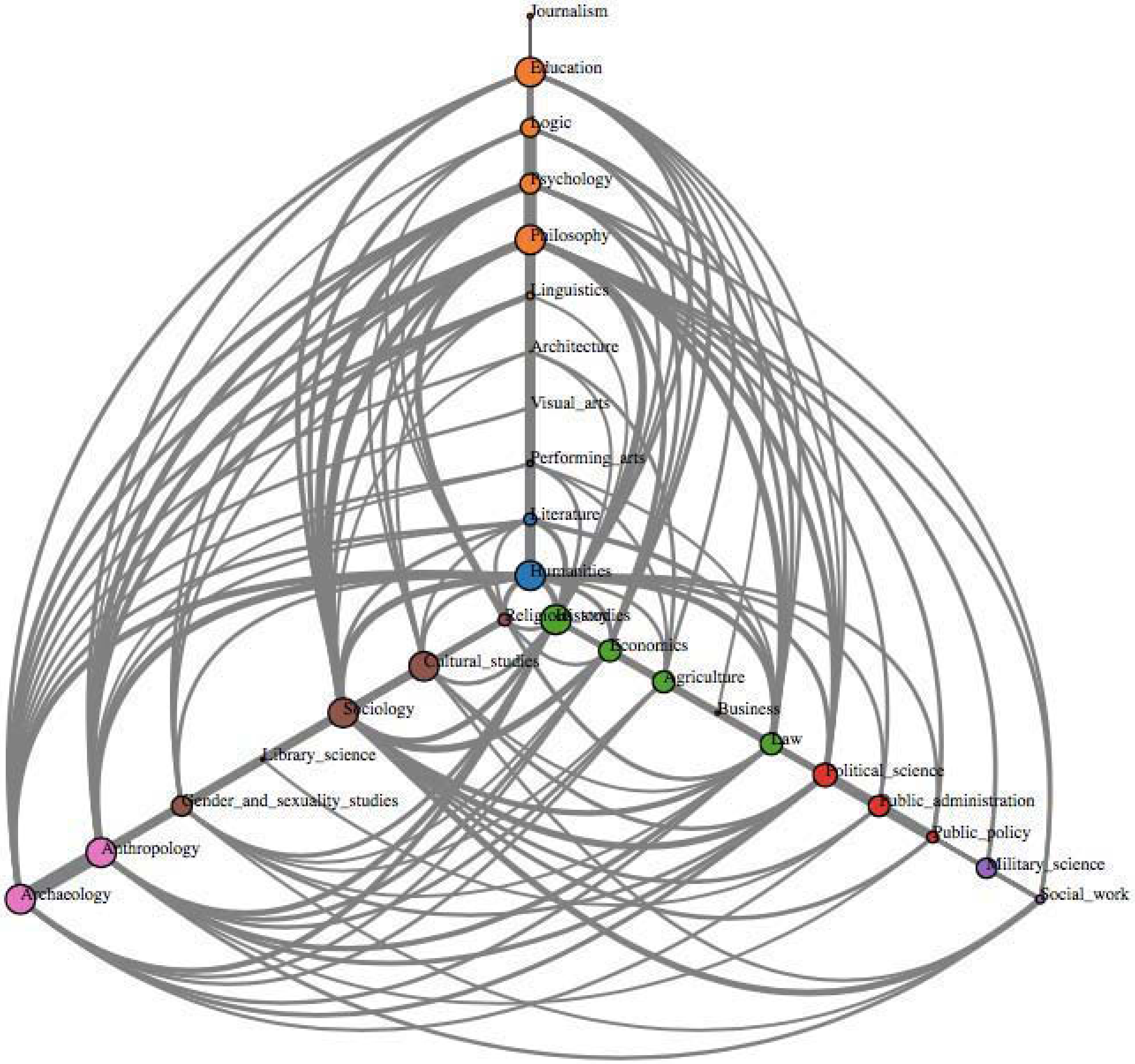


Figure A.35: Hive plot of subclustering of Canonical++ dataset, top cluster, Weight greater than or equal to 5



Appendix B

Tables

Table B.1: Summary of factors that introduce error into word occurrence based estimates of conceptual divergence.

Factor	**Description**	**Δ Conceptual Divergence**
Deterritorialization	Appropriation of a concept by a neighboring group causes a shift in meaning without a corresponding shift in spelling	Underestimate
Metaphor	Non-literal use of term	Underestimate
Weakening	A term has fewer restrictions placed on it in a context	Underestimate
Appropriation	Token is exploited by an outside group for their own ends	Underestimate
Typo	Typographical errors	Overestimate
Reterritorialization	Increase or shift in connotation following deterritorialization of a term/ redefinition without corresponding shift in spelling	Underestimate
Metonymy	An attribute of something is used to refer to that thing	Overestimate
Shortening/Ellipsis	Reduction of a full spelling. Ex/ periodical paper' to 'periodical'	Overestimate
Synecdoche	A part of something is used to refer to that thing	Overestimate
Group Divergence	Concept use within two groups differs without drastic change in lexicon	Underestimate
Root Word Similarity	Two words have similarly spelled roots, or the same root	Underestimate

Table B.2: Summary of differences between number of hyperlinked and whole page keywords in Canonical+ dataset.

Disciplines	Hyperlinked Keywords	Whole Page Keywords	Difference
Cosmology	311	1242	931
Meterology	349	1942	1593
Logic	351	1818	1467
Biochemistry	355	1382	1027
Mathematics	435	1719	1284
Biology	464	1810	1346
Archaeology	510	2689	2179
Chemistry	531	1747	1216
Geology	532	1966	1434
Philosophy	679	3095	2416
Quantum_mechanics	738	2184	1446
Psychology	760	4016	3256
Physics	805	1814	1009
Sociology	814	3195	2381
Anthropology	1081	2977	1896

Table B.3: Summary of the differences in the fraction of informative keywords for the Canonical+ dataset when hyperlinked and whole page text is used.

Disciplines	inffr(Canonical+)	inffr(Canonical+WholePage)	Δ
Anthropology	0.321626617	0.646210597	0.32458398
Archaeology	0.390196078	0.69910847	0.308912391
Biochemistry	0.211267606	0.722865412	0.511597807
Biology	0.321888412	0.73814774	0.416259328
Chemistry	0.338432122	0.781017724	0.442585602
Cosmology	0.321656051	0.753968254	0.432312203
Geology	0.127819549	0.671573604	0.543754055
Logic	0.455840456	0.821428571	0.365588116
Mathematics	0.337931034	0.799537839	0.461606805
Meterology	0.171428571	0.672230653	0.500802081
Philosophy	0.393603936	0.71319797	0.319594034
Physics	0.61242236	0.843148046	0.230725686
Psychology	0.171052632	0.639103362	0.468050731
Quantum_mechanics	0.516260163	0.788197621	0.271937459
Sociology	0.224815725	0.713347921	0.488532196
Average	**0.327749421**	**0.733538919**	**0.405789498**

Table B.4: Normalized and Non-Normalized betweenness centrality measures for the Canonical+ dataset using only hyperlinked keywords in a G_t graph. A graph for the full text extraction $G_t^{wholepage}$ was also constructed. Values of 0 indicate that discipline was never included in the shortest path between any other pair of disciplines, i.e. $G_t^{wholepage}$ is a rhizome.

Disciplines	$g(v_i)$	$g_{normalized}(v_i)$	$g(v_i)$ for $G_t^{wholepage}$
Biology	0.928616697	1	0.0
Chemistry	0.752338958	0.81017169	0.0
Physics	0.632818977	0.681464138	0.0
Psychology	0.360443016	0.38815048	0.0
Mathematics	0.342500582	0.3688288	0.0
Quantum_mechanics	0.341631702	0.367893129	0.0
Archaeology	0.251035747	0.270333011	0.0
Anthropology	0.235992821	0.254133726	0.0
Meterology	0.088453801	0.095253296	0.0
Geology	0.066167698	0.071254047	0.0
Logic	0	0	0.0
Cosmology	0	0	0.0
Biochemistry	0	0	0.0
Sociology	0	0	0.0
Philosophy	0	0	0.0

Table B.5: Top 3 linear hierarchical ordering ranked by R^2. Discipline acting as pole appears uppermost and boldface. R^2 and *pavlue* for each regression analysis listed above each column.

1st	2nd	3rd
$R^2 = 0.103$, $p = 0.016$	$R^2 = 0.087$, $p = 0.028$	$R^2 = 0.072$, $p = 0.046$
Sociology	**Physics**	**Biochemistry**
Psychology	Quantum_mechanics	Biology
Philosophy	Cosmology	Chemistry
Anthropology	Chemistry	Physics
Physics	Philosophy	Quantum_mechanics
Biology	Psychology	Anthropology
Cosmology	Biology	Geology
Quantum_mechanics	Sociology	Psychology
Chemistry	Geology	Sociology
Geology	Anthropology	Philosophy
Biochemistry	Biochemistry	Cosmology

Table B.6: Comparison of Canonical dataset and highly central nodes in Canonical++. When page does not appear in Canonical++ it is marked NA, and when $g(v_i) \gg 0$ it is marked with a series of Xs.

Canonical Dataset	$g(v_i) \gg 0$ for $v_i \in$ Canonical++
Cosmology	NA
Geology	NA
Philosophy	XXXXXXX
Chemistry	XXXXXXX
Biochemistry	XXXXXXX
Sociology	Sociology
Psychology	Psychology
Anthropology	Anthropology
Physics	Physics
Biology	Biology
Quantum_Mechanics	Quantum_Mechanics
	Meteorology
	Mathematics
	Computer_Science
	Agriculture
	Geography

Bibliography

[Bandelt and Dress 1992] Bandelt, H. J., & Dress, A. W. (1992). Split decomposition: a new and useful approach to phylogenetic analysis of distance data. *Molecular Phylogenetics and Evolution,* **1**(3) 242-252.

[Bapteste et al 2012] Bapteste, E., Lopez, P., Bouchard, F., Baquero, F., McInerney, J. O., & Burian, R. M. (2012). Evolutionary analyses of non-genealogical bonds produced by introgressive descent. *Proceedings of the National Academy of Sciences*, **109**(45) 18266-18272.

[Bird et al. 2009] Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python. *O'Reilly Media.*

[Bostock et al. 2011] Bostock, M., Ogievetsky, V., & Heer, J. (2011). D^3 data-driven documents. *Visualization and Computer Graphics, IEEE Transactions on,* **17**(12), 2301-2309.

[Brewer 2006] Brewer, D. (2006). The discourse of enlightenment in eighteenth-century France: Diderot and the art of philosophizing. *Cambridge University Press*, (Vol. 42).

[Buckner et al. 2011] Buckner, C., Niepert, M., & Allen, C. (2011). From encyclopedia to ontology: Toward dynamic representation of the discipline of philosophy. *Synthese,* **182**(2), 205-233.

[Callon et al. 1990] Callon, M., Courtial, J. P., & Laville, F. (1990) Co-word Analysis as Tool for Describing the Network of Interactions Between Basic and Technological Research: the Case of Polymer Science. *Scientometrics*, **22**(1) 155-204

[Compt 1835] Comte A (1835) Cours de philosophie positive: Borrani et Droz.

[Csardi and Nepusz 2006] Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695. http://igraph.org

[Darnton 2009] Darnton, R. (2009). The Business of Enlightenment: a publishing history of the Encyclopédie. *Harvard University Press*, 1775-1800.

[Deleuze and Guattari 1980] Deleuze, G., & Guattari, F. (1980). Capitalisme et Schizophrénie. *Paris,* (Vol. 2).

[Diderot and d'Alembert 1772] Diderot, D., & d'Alembert, J. L. R. (1772). Explication détaillée du système des connaissances humaines.

[Doolittle et al. 2014] Doolittle, W. F., Brunet, T. D., Linquist, S., & Gregory, T. R. (2014). Distinguishing between “function” and “effect” in genome biology. *Genome Biology and Evolution*, **6**(5), 1234-1237.

[Dopanzo et al. 1993] Dopazo, J., Dress, A., & Von Haeseler, A. (1993). Split decomposition: A technique to analyze viral evolution. *Proceedings of the National Academy of Sciences*, **90**(21), 10320-10324.

[Dupré 1983] Dupré J. (1983) The Disunity of Science. Mind Association. *Oxford University Press*, **92**(365) 321-346

[EigenFactor 2015] EigenFactor, (2015). About the EigenFactor Project. *http://www.eigenfactor.org/about.php* Accessed: September 8th, 2015.

[Encyclopædia Britannica Inc. 2006] Encyclopædia Britannica, Inc. (2006) Fatally Flawed: Refuting the recent study on encyclopedic accuracy by the journal Nature.

[Fanelli and Glänzel 2013] Fanelli, D., Glänzel, W. (2013). Bibliometric Evidence for a Hierarchy of Science. *PLoS ONE* **8**(6): e66938. Dio:10.1371/Journal.pone.0066938

[Foucault 1969] Foucault, M. (1969). L’Archéologie du Savoir. *Éditions Gallimard*

[Foucault 1970] Foucault, M., (1970). Les Mots et les Choses: Une Archéologie des Sciences Humaines. *Pantheon Books.*

[Frege 1879] Frege, G. (1879). Begriffsschrift: eine der arithmetischen nachgebildete Formelsprache des reinen Denkens. *Halle*

[Giles 2005] Giles, J. (2005) Internet Encyclopedias Go Head to Head. Nature **438**, 900-901 (15 December 2005) — doi:10.1038/438900a; Published online 14 December 2005

[Girvan and Newman 2002] Girvan, M., & Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences,* **99**(12), 7821-7826.

[Heggarty 2006] Heggarty, P. (2006). Interdisciplinary indiscipline? Can phylogenetic methods meaningfully be applied to language data and to dating language. *Phylogenetic Methods and the Prehistory of Languages*, 183.

[Hoskin and Macve 1986] Hoskin, K. W., & Macve, R. H. (1986). Accounting and the examination: a genealogy of disciplinary power. *Accounting, Organizations and Society,* **11**(2), 105-136.

[Huson and Bryant 2006] Huson, D. H., & Bryant, D. (2006). Application of Phylogenetic Networks in Evolutionary Studies, *Mol. Biol. Evol.*, **23**(2) 254-267.

[Huson and Bryant 2006] Huson, D. H., & Bryant, D. (2006). User Manual for SplitsTree4 V4.6.

[Keller 1991] Keller, E. F. (1991). Fractured Images of Science, Language, and Power: A Postmodern Optic, or Just Bad Eyesight? *Poetics Today*, 227-243.

[Kitcher 1984] Kitcher, P. (1984). 1953 and All That. A Tale of Two Sciences. *The Philosophical Review*, **93**(3) 335-373.

[Knobe et al. 2008] Knobe, J., & Nichols, S. (2008). An experimental philosophy manifesto. *Experimental philosophy*, 3-14.

[Krzywinski et al. 2012] Krzywinski, M., Birol, I., Jones, S. J., & Marra, M. A. (2012). Hive plotsrational approach to visualizing networks. *Briefings in Bioinformatics,* **13**(5), 627-644.

[Lenoir 1993] Lenoir, T. (1993). The discipline of nature and the nature of disciplines. *Knowledges: Historical and critical studies in disciplinarity*, 70-102.

[Mantel 1967] Mantle, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research* **27**(2) 209-220. PMID 6018555.

[Masucci et al. 2011] Masucci A. P., Kalampokis, A., Eguíluz, V. M., Hernández-García, E. (2011) Wikipedia Information Flow Analysis Reveals the Scale-Free Architecture of the Semantic Space. *PLoS ONE* **6**(2): e17333. doi:10.1371/journal.pone.0017333

[Mayr 1982] Mayr, E. (1982). The growth of biological thought: diversity, evolution, and inheritance. *Harvard University Press.*, 205.

[Morrison 2014] Morrison, D. A. (2014). Is the Tree of Life the Best Metaphor, Model, or Heuristic for Phylogenetics? *Syst. Biol.* **63**(4) 628638.

[Nietzsche 1956] Nietzsche, F. W. (1956). The Birth of Tragedy and the Genealogy of Morals. *Anchor Books.* (Vol. 81).

[Newman 2006] Newman, M. E. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences,* **103**(23), 8577-8582.

[Nielsen 2007] Nielsen, F. Å. (2007). Scientific citations in Wikipedia. *arXiv preprint*, 0705.2106.

[Qin 1999] Qin, H. (1999) Knowledge Discovery Thought Co-word Analysis. *Library Trends*, **48**(1) 133-159.

[Rambaut et al. 2001] Rambaut, A., Robertson, D. L., Pybus, O. G., Peeters, M., & Holmes, E. C. (2001). Human immunodeficiency virus: phylogeny and the origin of HIV-1. *Nature,* **410**(6832) 1047-1048.

[Saitou and Nei 1987] Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution,* **4**(4) 406-425.

[de Saussure 1983] de Saussure, F. (1983) Course in General Linguistics *Eds. Charles Bally and Albert Sechehaye. Trans. Roy Harris.* La Salle, Illinois: Open Court.

[SEP] *The Stanford Encyclopedia of Philosophy.* Principal Editor: Edward N. Zalta, (http://plato.stanford.edu/info.html, accessed: 2015/08/27)., The Metaphysics Research Lab, Center for the Study of Language and Information, Stanford University, Stanford, CA 94305-4115.

[Sokal and Michener 1958] Sokal, R., & Michener, C. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin* **38**: 14091438.

[Sowa 2006] Sowa, J. F. (2006). Semantic networks. *Encyclopedia of Cognitive Science.*

[Thagard, 1993] Thagard, P. (1993). Computational philosophy of science. *MIT press.*

[Woese and Fox 1977] Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences,* **74**(11), 5088-5090.

[Zlatić et al. 2006] V. Zlatić, M. Božičević, H. Štefančić, and M. Domazet (2006) Wikipedias: Collaborative web-based encyclopedias as complex networks.*Physical Review E*, **74**(1), 016115.