ScuDo

Scuola di Dottorato ~ Doctoral School

WHAT YOU ARE, TAKES YOU FAR

Doctoral Dissertation

Doctoral Program in Applied Mathematics (29^{th}cycle)

Human exploration of complex knowledge spaces

By

Giovanna Chiara Rodi

******

Supervisor(s):

Prof. V. Loreto, Supervisor

Prof. L. Rondoni, Supervisor

Dott.ssa F. Tria, Co-Supervisor

Doctoral Examination Committee:

Prof. A. Flammini, Referee, Indiana University, USA

Dott. B. Gonçalves, Referee, New York University, USA

Prof. M. Caselle, University of Turin, Italy

Prof. L. Mesin, Politecnico di Torino, Italy

Prof. F. Vaccarino, Politecnico di Torino, Italy

Politecnico di Torino

2017

Declaration

I hereby declare that, the contents and organization of this dissertation constitute my own original work and does not compromise in any way the rights of third parties, including those relating to the security of personal data.

Giovanna Chiara Rodi
2017

* This dissertation is presented in partial fulfillment of the requirements for **Ph.D. degree** in the Graduate School of Politecnico di Torino (ScuDo).

Abstract

Driven by need or curiosity, as humans we constantly act as information seekers. Whenever we work, study, play, we naturally look for information in spaces where pieces of our knowledge and culture are linked through semantic and logic relations. Nowadays, far from being just an abstraction, these *information spaces* are complex structures widespread and easily accessible via techno-systems: from the whole World Wide Web to the paramount example of Wikipedia. They are all *information networks*.

How we move on these networks and how our learning experience could be made more efficient while exploring them are the key questions investigated in the present thesis. To this end concepts, tools and models from graph theory and complex systems analysis are borrowed to combine empirical observations of real behaviours of users in knowledge spaces with some theoretical findings of cognitive science research.

It is investigated how the *knowledge space* structure can affect its own exploration in learning-type tasks, and how users do typically explore the information networks, when looking for information or following some learning paths. The research approach followed is exploratory and moves along three main lines of research.

Enlarging a previous work in algorithmic education, the first contribution focuses on the topological properties of the information network and how they affect the *efficiency* of a simulated learning exploration. To this end a general class of algorithms is introduced that, standing on well-established findings on educational scheduling, captures some of the behaviours of an individual moving in a knowledge space while learning. In exploring this space, learners move along connections, periodically revisiting some concepts, and sometimes jumping on very distant ones. To investigate the effect of networked information structures on the dynamics, both synthetic and real-world graphs are considered, such as subsections of Wikipedia and

word-association graphs. The existence is revealed of optimal topological structures for the defined learning dynamics. They feature small-world and scale-free properties with a balance between the number of hubs and of the least connected items. Surprisingly the real-world networks analysed turn out to be close to optimality.

To uncover the role of semantic content of the bit of information to be learned in a information-seeking tasks, empirical data on user traffic logs in the Wikipedia system are then considered. From these, and by means of first-order Markov chain models, some users paths over the encyclopaedia can be simulated and treated as proxies for the real paths. They are then analysed in an abstract semantic level, by mapping the individual pages into points of a semantic reduced space. Recurrent patterns along the walks emerge, even more evident when contrasted with paths originated in information-seeking goal oriented games, thus providing some hints about the unconstrained navigation of users while seeking for information.

Still, different systems need to be considered to evaluate longer and more constrained and structured learning dynamics. This is the focus of the third line of investigation, in which learning paths are extracted from advances scientific textbooks and treated as they were walks suggested by their authors throughout an underlying knowledge space.

Strategies to extract the paths from the textbooks are proposed, and some preliminary results are discussed on their statistical properties. Moreover, by taking advantages of the Wikipedia information network, the Kauffman theory of *adjacent possible* is formalized in a learning context, thus introducing the *adjacent learnable* to refer to the part of the knowledge space explorable by the reader as she learns new concepts by following the suggested learning path. Along this perspective, the paths are analysed as particular realizations of the knowledge space explorations, thus allowing to quantitatively contrast different approaches to education.

Contents

Chapter 1

Introduction

Need or just curiosity pushes us, as humans, to continuously seek for information. In any activity, from work to play, we learn, retrieve old knowledges and simply forget, while the entire environment around us ceaselessly proposes new stimuli. In this sea of flowing information, our greatest effort is not to get lost, rather find our path through it.

Indeed, as soon as we look for any piece of information, we walk in a space. It could be individual – the space of our knowledges, memories and personal associations – or collective, such as the cultural heritage, as it results from the human evolution. These spaces are complex, evolving structures where pieces of our knowledge and culture are interlinked through semantic and logic relations. In these spaces we shape our ways, by moving from a bit of information to another, maybe wandering, until the temporary target is reached. More intriguingly, as soon as novelties are discovered, entire new possibilities enrich our personal space, thus becoming available for learning [68].

Far from being just an abstraction, these evolving information spaces are widespread, from the whole World Wide Web to the paramount example of Wikipedia [27, 118, 62], from word-association graphs [49, 55] to ontologies and taxonomies. Moreover, accessing them has become easier for more and more people, thanks to freely accessible internet and progresses in technologies. Along with this, the explosion since 2012 of Massive Open Online Courses (MOOCs) witnesses the exponential growth in the demand for access to education, as it is similarly done by the recent success of web platforms and applications designed for learning, e.g., Anki [40] or Duolingo [1].

Thus, the *knowledge space* is more accessible, but still it continuously enlarges and more and more people claim their right to learn and contribute to the collective knowledge. In this scenario, no enough teachers can account for the world-based demand for education. Still, technologies can fill the gap, for example by proving the society with innovative tools for learning, maybe tailored to the needs of every distinct self-learner [89].

To this end, the needed novel educational software should enable learners to efficiently define their proper ways in the information spaces, as modern global positioning systems allow human beings to locate themselves and find their way in the physical space. Thus, the very first challenge is to understand how the knowledge spaces are shaped, and how they should be in order to improve our experience as information seeker. Moreover, understanding how we behave as learners [83, 57], and if any common pattern exists in our behaviours, is crucial to rethought the classical educational schemes, not ready to account for the *complexity* of the challenge.

Indeed, not only is the cognitive process of *learning* deeply complex. Learning is a matter of brain interconnections, of memory, of elaboration of information, and still before of perception of the world around us. This is of course a complex process that filters out enormous amounts of data and flags potentially relevant information to allow the individual to navigate the world as function of her needs and the environment's affordances [48, 66, 99].

Yet, another source of complexity should be considered, namely that of the *knowledge space* in which we move while looking for information. The space can be thought as the co-evolving overlapping of individual and collective spaces, their structures reflecting the complexity of the phenomena originating it, such as of our brain evolution and also of the continuous and evolving relationships between us, as for example in language. Indeed, even when we consider some very well defined entities, such as the thesaurus of word of a particular language, still this entities feature properties of complexities.

As a consequence, if we want to tackle the complexity of the learning challenge, first of all the structure of the space over which we move needs to be formally represented, to encompass its richness, heterogeneity and dynamical nature. Indeed, every information space can be though as a bunch of pieces of information, interconnected by logical, semantic, phonetic, spatial, temporal, linguistic relations, just to draw few.

Complex network analysis, and more in general complex systems theory, provides the proper formal and quantitative framework to investigate it.

Networks allow to represent associations, graph theory and statistical analysis are the tools to investigate the space properties, maybe recurrent in diverse systems. Moreover, any information-seeking task can be modelled as an exploration of the networked space, thus allowing to investigate both the role of the topology in hindering or enhancing the exploration and the common ways the exploration is performed.

In some systems, the networked representation spontaneously emerges. That is the case of all hyperlinked systems, such as the World Wide Web and, even more related to the picture of an *networked knowledge space*, Wikipedia [7], often considered as a primary source of well-established and reliable information. It represents a huge system of pieces of knowledge, continuously created, updated and modified by users, as well as the networked structure in which they are embedded. In this *knowledge space*, our walks from article to article can be effectively thought of as *learning* paths [88], either oriented towards very specific pieces of information or moved by curiosity. Comprehending how we act in such a space could be key to design new and more effective learning strategies as well as to improve the actual systems and platforms designed for knowledge search.

That is the proper aim of the present thesis. Within the complex systems framework, it is investigated how the *knowledge space* structure can affect its own exploration in learning-type tasks, and how users do typically explore the accessible information networks, when they look for information or follow some learning paths. The research approach followed is exploratory. Far from giving exact, quantitative results, three main lines of research are defined, along with novel concepts and representations to deal with available data and previous results on cognition, learning and information networks.

The first line of investigation [96] aims at enlarging a previous work by Novikoff et al. [88] on algorithmic education, i.e., algorithms which are able to propose to a learner the *optimal* scheduling of the study sessions of any bunch of items to be learned. Optimal, in their work, refers to a temporal notion. Indeed, there are results from cognitive research [39] stating that some temporal constraints should be considered while balancing the introduction of new materials and the repetition of old, already acquired, ones, in order to minimize forgetting. Novikoff et al. formalized

mathematically this problem and developed some models for the generation of learning schedules that would yield to lifelong learning or cramming, without any forgetting during the time considered. However, in their scheme, no correlation between the units to be learned is taken into account, while empirical results suggests that they affect learning [18].

The work here proposed moves along this direction, by using a complex network representation of the units to be acquired and their interconnections. Their possible effects in learning are discussed and modelled, based on previous literature in cognition and word learning. With all this, a stochastic algorithm is devised to order the introduction and reviews of the interconnected materials, while still satisfying constraints on timing. Far from being a truly educational software, or a model of actual learning dynamics, it allows to investigate which topological properties are key to boost the efficiency of exploration, and if these crucial properties are found in information networks based on real systems.

With the starting aim of further enriching the proposed model with empirical observations of how information-seekers tend to explore real information spaces, the research effort has moved to a different track. The investigation of the dataset of clickstream of users in Wikipedia is the research object of the second contribution presented in this thesis. From this dataset, the possible paths followed by Wikipedia readers are simulated and analysed to identify possible statistical patterns and regularities. To this end, a novel abstract representation of the Wikipedia articles is devised, by which any article is mapped into a multidimensional space of broad general topics, thus synthesising their topical content. The proposed topical mapping is based on the category systems of Wikipedia, thus relying on the collective and aggregate perception of knowledge, as it emerges in the online encyclopaedia and its categorical structure.

The dataset and the scheme defined allow to distinguish between different browsing strategies, according to the task motivating the navigation and the source from which the user enters Wikipedia, thus posing the presented contribution [97] close to other researches on semantic driven navigation of user in information networks. Still, the analysis done on the exploration habits of Wikipedia readers is based on simulated paths of users across pages in the encyclopaedia. The overcome this problem, and rather focusing on true *learning paths*, some textbooks of advanced scientific

content are also considered, thus defining the third line of research presented in this manuscript.

Indeed, any textbook could be though as a particular path across the knowledge space, namely the path suggested by the textbook author. The order with which different concepts are presented is assumed to be a result of both preliminarity constraints and style of writing and teaching. The writer chooses how and when to present the educational material as well as when reviewing old one. However, she carries the reader in the exploration of an (partially) unknown – conceptually, for the reader – area of the knowledge space. Indeed, while we read a course book of follow some lectures, any concepts introduced pave the way for others, thus uncovering many potential new paths of learning and exploration. In this sense, it is here realized the idea of *adjacent possible* proposed by Kauffman [68] in a biological framework. Every textbook is indeed one possible walk in the known space of the author, while the adjacent possible of the reader, namely what has become available for learning, enlarges as she comes across the book.

With this perspective, the textbooks are analysed, firstly by looking for the supposed underlying knowledge space. To this end, the classical level of analysis in quantitative linguistics, namely words and n-grams, is left for a more conceptual level of description. The texts are evaluated as streams of *semantic units*, extracted from each sentences. In particular the choice of using the external software TAGME [44] to tag each sentence into pages of Wikipedia allows for a direct mapping of each textbook into a path of exploration of the Wikipedia graph. Preliminary results are reported on this approach to the study of suggested learning paths, which could in future lead to the study of very different educational tools and the underlying teaching strategies.

Before entering into the details of the presented lines of investigations, brief reviews are reported on the contributions of complex network theory to the analysis of information systems, such as language, and of information-retrieval tasks, in memory as well as in actual information hyperlinked systems.

Chapter 2

Semantic network

For the last decades, concepts, tools and methods borrowed from statistical physics have been widely applied to the study of diverse, interdisciplinary phenomena, all characterized by features of *complexity*. Typically, with *complexity* it is denoted the spontaneous appearing of macroscopic properties, from microscopic interactions of their constituents. Notably, common regularities, features and quantitative behaviours are recovered over several different types of systems, from biological to socio-economics ones. Thus, the scientists aiming at formalizing so diverse phenomena have focused on the structural properties of the systems, their "*organization, their pattern and form*", rather than their single constituents [25].

To this end, networks have proven to be the proper structures to represent a large variety of complex systems [9, 85], as well as where to reproduce dynamical processes of interactions. In the complex system perspective, a network is a graph of nodes connected by edges, which can abstractly represent any type of interaction or relationship between the nodes. Because of this, and since the complex network approach allows to tackle the large amount of data from techno-social systems, which are nowadays available, it has resulted in a extremely versatile mathematical tool.

The novel quantitative perspective offered by complex network theory has deeply influenced the research in many different domains, like language and cognition sciences. In this area, complex networks theory has contributed with its own more abstract, coarse-grained approach, thus abroad addressing questions about structure and universal properties of linguistic systems as well as of cognitive processes, with respect to the traditional methods of linguistics and psycholinguistics.

Besides language, other collective information systems are widely represented through complex networks, and, as well as linguistic systems, they are also shown to present statistical signatures of complexity. That is the case of the World Wide Web and of its smaller subgraph Wikipedia.

Indeed, the mentioned systems are by their own nature networks, where pages are connected through hyperlinks. As for language, the structure and properties of such information systems have been largely studied. Moreover, thanks to the vast amount of available data, also the users' behaviour on them has been investigated, by looking for patterns and models to explain the empirical observations.

In the present chapter some essential literature is reviewed in order to define the framework of the work presented in the following. In particular, in the first section some results of the complex networks theory approach to language and cognition are surveyed. Far broad reviews on the topic can be found in the works of Borge-Holthoefer et al. [22] and Baronchelli et al. [15], here considered as main references. Then, results are presented on research carried on the structures of and the user behaviours on complex *information graphs*.

2.1 Semantics, networks and cognition

2.1.1 Linguistic networks

As already pointed out, graph theory and complex networks advances have allowed to gain a novel perspective in the study of human language [18]. Indeed, in the complex networks framework, the complexity of language structure, its generation and evolution can be tackled in a formal representation [103], thus overcoming the classical linguistic approach, more focused on a static analysis of language.

The first issue of a complex network approach to any system is of course the identification of the elements constituting the networks and of the type of associations to depict. Language networks can be built upon different types of human data, usually extracted from empirical observations of language use as in free association tasks, in corpora of written texts, in ad-hoc designed experiments. In all cases, several different linguistic aspects can be selected, such as words phoneme, semantics, morphology etc. Still, whatever the type of relationships between words, they can be

depicted by means of a graph, whose structure and topology reflect the complexity of the words connections.

The first contribution of complex network science has indeed focused on the structure of the various language graphs. Interestingly, different language networks share statistical features at a global level. Why these common signatures appear, and what they reflect about the key properties of the language organization and evolution, both at individual and collective level, have become the research questions of the networks scientists interested in language [103]. The object of analysis has then shifted from the mere structural properties of language networks to the processes leading to their formation as well taking place on them, thus putting the complex network approach closer to the cognitive one [22, 18].

Statistical signatures of language complexity

One type of network widely used to investigate words organization into language is the *co-occurrences* networks. In them, starting from a corpus of texts, words are connected when appearing close in a text, up to a distance to be chosen. Changing the distance allows indeed to take into account different types of correlations between words [45, 21]. The work of Ferrer-i-Cancho and Solé [45] is the first broad investigation of co-occurrences networks of English language. They recover the presence of features of complexity common to several other social and biological networks, namely the small-world effect [110] and a power-law degree distribution.

The former, firstly observed by Watts and Strogatz [110], consists in the co-presence of two different topological properties. Networks are small-world if the average shortest path between any pairs of node is *small*, i.e., similar to the one found in a random graph with same order and average connectivity. Moreover, they have to display a transitivity higher than in a comparable random graph. With transitivity the average clustering coefficient of the network is referred to, which measures how much the network is clustered, i.e., how many pairs of neighbours of a node are also neighbours of each other, on average. Together with these properties, the power-law scaling of the degree distribution, found in the co-occurrences graphs, is another common signature of complex systems and typically associated to a particular generation rule, namely the preferential attachment [14]. In particular for the co-occurrences graph, it is found that two classes of words can be distinguished,

with different co-occurrences frequency profile, thus resulting in a double slope of the degree distribution [45].

The above mentioned properties, small-world effect and power-law scaling of degree distribution, are actually recovered also in other networked representations of human language [103]. For instance, Steyvers's et al. [105] analyse different types of linguistic graphs, based on data collected in a free word association experiments, and on two diverse thesauri of words. Although the three graphs represent different semantic relations between words as originated from three diverse processes, the resulting semantic networks share the above mentioned statistical properties, typical of scale-free complex structures.

The fact that statistical signatures of complexities are ubiquitous in different semantic representation of language suggests that they could actually have a cognitive role, i.e., they could reflect properties of human cognitive processes [18]. For instance, small-world properties in networks have been proven to lead to efficient and robust retrieval of information [31]. Research has then focused on possible generative end evolutionary models for the linguistics networks, with an effort to put them in the right cognitive framework.

Lexicon generation and language use

The recurring small-world, scale free topologies recovered in various language networked representations have suggested that a underlying mechanism of language acquisition should exist. To uncover this, different models have been presented and tested against real data. Among the others, the generative model presented by Stevyers and Tennenbaum [105] provides one "*testable hypothesis on how structure in language might emerge during the process of acquisition*" [18]. Indeed, to account for the observed scale-free patterns and high clustering, the authors propose a modification of the preferential attachment model [14]. In their augmented version, any new node entering the network during acquisition *differentiates* an existing one, chosen with probability proportional to its actual degree, as in the preferential attachment rule. The introduced differentiation consists in the fact that any novel node is connected to a subset of neighbours of the chosen existing one, as to differentiate its meaning/context. This mechanism is proven to generate networks with properties comparable to what found analysing the adult semantic network. Moreover it provides an explanation for another empirical results, namely the positive correlation

between word age of acquisition, their frequencies and their connectivity profile in semantic networks.

Along the same direction, the works of Hills et al. [60, 59] and of Beckage [17] investigate the *developmental networks*, namely the language networks as they evolve in children. These are based on data registered by parents about production of words in their children during early months of language acquisition. Connections between learned words are then assumed by referring to other corpus or thesaurus, as associative of features norms.

By only looking at the structural differences between development network of individual children, atypical patterns of language development can be identified, thus allowing to predict typical and late talkers [17]. On the aggregate dataset, typical (normative) networks of children with different ages can be obtained, and used to test language acquisition model.

In [60, 59] three mechanisms of word acquisition are tested. The first one is the preferential attachment, according to which words are early acquired if they would connect to the *hubs* of the known network at each stage, i.e., to the known words which also are the mostly connected. The other two mechanisms are the novel contributions of the authors, namely the *preferential acquisition* and the *lure of associates*. With the former, words are earlier introduced if they are more connected in the learning environment, i.e., to the most connected words in the adult semantic network. According to the *the lure of associates*, instead, words with more connections with the known words are early acquired.

While all three models fail to account for the development of feature networks, the preferential acquisition model and the lure of associates outperform the preferential attachment in predicting the evolution of association networks, thus uncovering the role that *contextual diversity* [59] can have in enhancing words learning in children. Indeed, the more diverse is the context of a word in the adult semantic network, the more frequently that word can be used with other words in the children learning environment, thus correlating with its observed early age of acquisition.

As emphasized by Beckage [18], the results in [60, 59] point out the relevance of language representation used to gain insight into cognitive processes. In this perspective, a very recent contribution has shown the promising power of a novel development of complex network theory, namely multilayer networks. Indeed, a comprehensive network representation of language has been proposed by Stella et

al. [104]. The authors introduce the multiplex lexical networks as a unique mathematical structure which absorbs 4 different interconnected layers, corresponding to different networked language representations (associative, co-occurrences, features and phonological networks). On this structure, where early learning acquisition is investigated, the superiority of the novel, multiplex approach is shown, for instance in detecting different stages of the language development in children.

2.1.2 Information retrieval and memory

When knowledge, as words or concepts, are represented by means of a network, then several cognitive processes may be modelled as dynamical processes placed on it. This could be the case of *information retrieval* tasks, i.e., when it is asked to find the way to a target piece of information, as well as of mental exploration [22], when instead from a stimulus the memory is activated as in a broad search without targets. Both types of tasks find in the network representation of memory the ideal structure where to be simulated as navigational processes, thus allowing to investigate how their efficiency if naturally affected by the network topology, i.e., by their own organization.

The milestone contributions in this type of approach on cognitive processes is found in the Quillian's idea of *semantic memory* and later extensions. In their work in 1969, Collins and Quillian propose a tree-like hierarchical model of the human semantic memory, i.e., the memory which includes "*properties of language storage and retrieval*" [18], according to the psychological definition. In their model, the semantic memory is represented so that concepts are nodes and connections represent class-inclusion relations. On this type of representation, the retrieval times for sentences of the type "A canary can fly" [32] are computed in a computer and then contrasted with real human retrieval times. The model over which the search process on the semantic network is based is called *spreading activation model*. It assumes that, from the input concepts, the search diffuses in parallel over the links, activating a tag on all the nodes visited, until a node already activated is met (in this case the path followed to connect the two concepts is evaluated and the information is possibly retrieved), or indefinitely, in the case of priming, as for mental navigation.

The constrained model proposed by Quillian was then extended in [31] to account for other empirical observations on cognitive tasks. The final proposed model of

semantic memory assumes a more general graph, where concepts can be not only nouns but also adjectives. Connections here are weighted with the aggregate semantic similarity of the nodes connected, that is how many properties are shared by them. Still, activation spreads as in a breadth-first diffusion from the stimulated inputs, decreasing with the distance thus activating other nodes with different strengths. This activation model is in the cited work tested to ease information retrieval, and indeed the authors shown that human performances in retrieval tasks can be effectively investigated as processes taking place on networks [18].

In the following decades, the very simple psychological mechanism suggested by Quillian and Collins has inspired further investigations on the possible role in cognitive tasks of structural properties of nodes in language networks. That the case, for example, of the possible effect of neighbourhood density of nodes in phonological networks on retrieval and word identification tasks [109], and, more in general of the connectivity profile in semantic networks. For instance, Griffiths et al. [56] propose and observe a novel correspondence between human memory organization and functioning and PageRank [90] algorithm. They focus on investigating *fluency*, namely how fast people can retrieve different information, e.g., *all words starting with B*. This type of retrieval processes is claimed to be very similar, computationally, to any Internet search engine, as Google. Indeed, the problem in both cases is to find the relevant items to answer a query, by looking in a large network of interconnected nodes – semantic concepts or Web pages. It follows the authors' idea to test, in a semantic memory network, the PageRank algorithm used by Google to compute the relevance of Web pages. Roughly speaking, the algorithm computes the centrality of each page by considering the importance as flowing across the links and pages, so that most relevant pages receive a major flux because they are connected to other highly relevant page, in a recursive definition. When the same algorithm is applied to a semantic network, the PageRank predictions outperformed other measures of word prominence in memory, like word usage frequency. It is worth to stress, as done by the authors, that the PageRank performance could indeed be explained by assuming the simple cognitive mechanism of *spreading activation* proposed by Collins and Loftus [31]. Indeed, in both cases the efficiency of the cognitive tasks depends on how information flows in the network [22], and thus finally on its topological properties.

2.2 Information graphs

2.2.1 Structure of information networks

The World Wide Web is an important example of information network, in which pages containing information are interconnected by hyperlinks. From its very beginning, research has focused on investigating its graph structure, as needed step both to gain insights into the generation of the system and to improve search engines and navigability.

On the Web structure, many different investigations have focused on both microscopic and macroscopic statistical properties. Since the very first researches [10, 71], small-world properties and scale-free profile in and out-degree distributions of pages were recovered.

In the first large-scale analysis, Broder et al. [24] confirmed the power-law distribution of in-degree and the large tail of the out-degree distribution, as well as for the scale-free distribution of strongly connected components. Furthermore, on a macroscopic scale, they proposed the bow-tie picture of the Web. Indeed, they observed a region of pages strongly connected (the CORE of the Web), a set of nodes not from which the COREs are reached, but not reachable from there (IN region) and a further region reachable from IN and CORE but do not directly connected to them (OUT region). Other set of nodes (TENDRILS and TUBES), are not part of the CORE, but still can be reached (or reach) IN (or OUT) nodes. This bow-tie structure was further observed in successive crawls, with the fraction of nodes belonging to each region depending on the crawl considered [38].

Interestingly, a similar bow-tie-like structure and connectivity profile was found by Capocci et al. [27] while investigating the Wikipedia graph. Indeed, as for the World Wide Web, the structure of pages connected by hyperlinks of the on-line encyclopaedia naturally suggests a graph-like representation, where directed edges interconnected the nodes/pages as hyperlinks do. The Wikipedia graph is continuously evolving, because the Wikipedia editors continuously add novel pages or, more frequently [27], novel edges between existing pages, thus possibly modifying the topological structure of the graph.

Still, other studies have pointed out the stability of its statistical microscopic properties, both across language versions [118] and across time [26]. By looking

at the temporal evolution of the Wikigraph, an evolving bow-tie structure is also recovered, the relative size of the CORE region enlarging as the OUT shrinks, and the graph becomes denser.

2.2.2 Browsing behaviours

The analysis of the structure of the hyperlinked network is key to investigate its navigability and the ease of information retrieval tasks by its users, exactly as for the semantic memory. Different models have been proposed to reproduce possible behaviours of users in information networks and how topology can affect traffic flows. In this sense, the paramount case is the PageRank algorithm of Brin and Page [23]. By assuming random surfers browsing the World Wide Web, their algorithm allows to rank pages in function of their centrality in the browsing activities.

Still, insights into the users browsing behaviours on information networks arrived as soon as data were available. From the very early studies of [64], attention has been focused on statistical regularities on how the WWW was explored, trying to quantitatively differentiate typical browsing strategies.

For instance, mainly based on server logs analysis of limited domains, studies have focused on characterizing the activity of web users, by observing cyclic regularities in their temporal patterns [53, 78]. Also the way users exploit the hyperlinked structure was object of analysis, revealing that teleportation via directly accessing pages of interest outperform navigation by following hyperlinks [78]. In [77] Meiss et al. contrasted individual patterns and aggregate patterns, recovering at the aggregate level the large tail distribution of site popularity (thus impeding any possible definition of *typical* traffic), as a result of log-normal distributions at the individual user levels.

Also regarding Wikipedia usage patterns, servers logs served as main dataset types for the early studies on the Wikipedia users. Ratkiewicz et al. [95] cross-correlated web requests outgoing the Indiana University over some months of browsing activities with article hits and other dataset to characterize traffic throughout the online encyclopaedia. Wikipedia in their analysis is observed to be a sink of traffic flowing into mostly from the same Wikipedia, from few referrers – typically search engines – and from empty referrer for around %9. Furthermore, by investigating the incoming flux from external sources and the outgoing flux originated in Wikidepia,

two predominant usage modes are recovered, namely *encyclopaedia* and *browsing*, in which respectively users tend to arrive from outside and remain in the Wiki, or to browse internally from page to page.

While being useful to understand local transitions between Wikipedia pages, log requests data do not provide complete information about individual user path. Despite the possibility of simulating paths from them, as later discussed in this thesis, other tools have been devised to obtain real navigation paths. For Wikipedia, games have been proposed in which users are asked to browse the encyclopaedia (or some reduced versions of it) while to fulfil the gaming task under some temporal or spatial constraints. *The Wiki Game* by Clemesha [30] and *Wikispeedia* [8, 112] are two notable examples.

In Wikispeedia players were asked to navigate on a reduced version of Wikipedia to go from a given starting page to a given target page always hopping on Wikipedia pages. Based on the paths gathered, West et al. reported [111] about hight efficiency of humans in identifying shortest paths and of emerging common navigation strategies. The Wikispeedia players early rely on hubs in the network, being they very common as first click. With this choice, they move from the first, assigned page, to a node where more options are typically available to push forward their task. While degree is crucial in the first phases on the games, after similarity becomes more important. Semantic similarity is computed by West et al. as a TF-IDF (text frequency-inverse document frequency) distance. With this they find that, after a few click, players get slowly semantically closer to their target, while also the semantic distance between consecutive pages visited diminishes.

Recently, the availability of large clickstream dataset extracted from the request logs of Wikipedia on a monthly bases for some months since January 2015 [115], has driven novel research on the issue of navigational behaviours of Wikipedia users. Indeed, these data report all requests of Wikipedia pages in the main namespace, i.e., articles and MainPage, performed by all users of the Desktop version during an entire month, thus providing crucial information to understand how navigability is affected by different features [72, 37]. That is the case of spacial/semantic features, concerning the position of the browsed hyperlinks in the article pages, of structural features related to the topological role of the pages in the Wikipedia graph or also semantic features. All these are observed in [37] to influence the *success* of a Wikipedia hyperlink, which is preferred if driving the reader towards the graph

periphery, semantically similar pages or if it placed in the top or left-hand side of the article.

Naturally, any kind of empirical observation about navigation preferences and patterns calls for navigational models which could account for them. Typically, it was done by properly steering the random surfer model, crucial in the navigation description because the main assumption of the PageRank [23] ranking algorithm.

Indeed, the classic model of a user who moves by following uniformly at random the hyperlinks while browsing information networks structure, sometimes jumping on the graph, has been investigated and modified, by introducing biases affecting how the link are selected at each step [50]. Still, whatever the rule according to which the next step is chosen, many models rely on a crucial property of the simulate random surfer, namely its *markovianity*.

With this, it is intended that the stochastic step of the modelled surfer is only affected by the particular transition probabilities from one page to another, regardless of its previous navigational history. Despite its extreme simplicity, this assumption has been recently proven to be statistically suitable to model the navigation of a system characterized by many states, as the pages of an information networks like Wikipedia. In particular, Singer et al. [102] tested varying order Markov chains models against navigational data from both goal-oriented and free tasks. Models are compared by means of different statistical model selection methods.

Given the sparsity of the available navigational data with respect to the high number of possible states, they conclude that memoryless Markov chains are the most suitable to model navigation on a page level. Indeed, they also report evidences that, as soon as navigation is considered between page *semantic features*, second order Markov chains models outperform the others. Thus, they recover that human navigation is indeed not memoryless, similar memory patterns emerging on a topical level of pages representation.

Chapter 3

Optimal dynamics on information networks

This chapter is dedicated to the description of the first line of research presented in this thesis, namely the investigation of the topological properties of information networks, on which an educational algorithm is simulated. This question is approached by starting from a previous, recent work on algorithmic education [88], in which the mathematical framework is defined of the timing issue in learning.

Indeed, as reported by a wide literature since 1885 with Ebbinghaus [39], it is known that timing is crucial in scheduling the study sessions of any material to boost the learning and minimize forgetting episodes. In [88] the problem of finding the proper balance between reviews and introductions of novel notions is addressed, and several scenario are discussed depending on the student performance, i.e. its rate of learning.

Here, the schema proposed in [88] is enlarged to account for interconnections between the units to be learned. Indeed, many empirical results suggest that associations between the material to be studied affect the learning process. The possible effects are formalized and taken into consideration while devising a class of educational algorithms. Such algorithms allow to explore, until complete coverage, the network of items to be learned by an abstract user, providing her with an ordered schedule of sessions in which either new material is proposed or old one is revised. With the devised algorithm, the role of the network structure embedding the item to-be-learned is investigated, by testing how different topological properties affects

the learning efficiency of the process simulated. To this end, both synthetic graph structures and real-world ones are tested, namely some subsections of the Wikipedia graph and the Human Brain Cloud network of free word associations [49].

The chapter is organized as follows. Firstly it is established the background of this work. Then the devised class of algorithms to explore an information network is presented, along with a discussion on the assumptions made to take into account possible effects from semantic associations between the units to be learned. The graphs used to test the algorithms as well as the real-based ones treated as proxies for real information networks are presented. Finally the results obtained when the educational processes are simulated on them are reported and discussed. The content, discussion and results hereafter discussed have been already published in [96].

3.1 Background

3.1.1 Timing issues in learning

In the last century, large discussions on learning and memory in the context of cognitive and psychological research have been conducted. Particular attention has been paid on the issue of allocating the study practices of items over time to gain the best learning performance, in terms of successful retrievals as well as long retention of the acquired knowledge. The milestone work is considered to be the one of Ebbinghaus in 1885 [39], in which he introduced the *spacing effect*. This finding refers to the notion that spreading the study sessions of any item over time makes its learning more durable than massing them in a short period, where the inter-study session intervals can be empty or filled with practices of other items. This effect is considered as "*one of the oldest and best documented phenomena in the history of learning and memory research*" [12]. Many different theories have been proposed to explain it in term of the psychological mechanisms involved. References are [34, 61] or the review by Dempster [36] where many experimental evidences of its validity can be also found.

Among all the possible inter-study intervals, research supports another finding on how to improve retention by optimally scheduling the reviews. It has been indeed reported [13] that the benefits gained by spacing are enhanced if, for each item, the intervals between its study practices expand with the reviews rather than remaining

fixed. This phenomenon is usually referred to with *lag effect* or *expanded retrieval*. A qualitative and quantitative review of the effect of these *distributed practice* can be found in [29]. By the way, many of the early experimental studies were related to immediate retention tasks, often involving exclusively paired-associate learning. Only recently more research has been focused on the effect of the distributed practices for long-term retention [12].

An example of practical formalization of these findings was already proposed by Pimsleur [94] who introduced and justified exponential expanding inter-study sessions. His approach is still now used as a valid language learning method. Several other efforts have been made in defining algorithms to compute the optimal times for practices [92] and in designing computer based learning systems, mainly based on flashcard (a famous example is the *SuperMemo* method and software package [114] or the most recent free software *Anki* [40]). Finally, another novel notable implementation of the spaced repetition technique, not based on flashcard procedure, is the free language-learning website *Duolingo* [1].

In addition to the results on the temporal constraints to make the learning more efficient, also the role of possible semantic connections between the materials to be studied are considered. This has been accomplished by referring to the the seminal spreading-activation theory for information retrieval [32, 31], and in particular to the assumption that, while learning, semantically related concepts could be primed or reinforced in memory. Further fundamental references are some results of previous research on the early words learning in toddlers [60, 59] or in second language learners, for which cognitive rather than linguistic associations seem to enhance the acquisition process [107], as already discussed in the previous chapter.

3.1.2 Algorithmic education

The spacing and lag effects defined are mathematically formalized as constraints in a learning schedule already in the cited work by Novikoff et al. [88]. They present a mathematical model to sequence educational material over time, in which the introductions and repetitions of abstract to-be-learned units are deterministically scheduled, thus defining a learning agenda in an *optimal way*. Their idea is to take the learner's needs and skills into account through two *spacing constraints* on the time useful for reviewing each item. For any given educational unit u_i, the temporal

distance between its k-th and $(k+1)$-th occurrence has to lie in the temporal window $[a_k, b_k]$, with $a_k \leq b_k$. Both the bounds are weakly increasing functions of the number of repetitions k. These two constraints, a_k and b_k, represent respectively the time before which any repetitions is useless for a better retention and the last time useful for reviewing before the item gets forgotten. The time steps interval $[a_k, b_k]$ is thus the temporal interval in which the *ideal* $(k+1)$-th review should occur to optimize the later retention.

More in details, the cited work focuses on two different educational goals: *cramming* and *lifelong learning*. The first one concerns an educational programme which aims at presenting and retaining a finite amount of units until a particular time. In contrast, the main goal of a lifelong learning is to schedule presentations and reviews in order to maximize the later retention, never forgetting anything, while the items to be learned grow without bounds. With reference to this last scenario, in their paper some algorithms are proposed to create agendas which allow to gain *infinite perfect learning* while satisfying different spacing constraints. Both constraints with $\{a_k,\ b_k\}$ exponential and polynomial in k are considered and the resulting learning rates are discussed. The function considered for quantifying the efficiency of the process is the introduction time function t_n, which accounts for the time when the n-th unit is introduced.

Among the several results presented, the authors show that schedules can be constructed for which both infinite perfect learning is achieved and t_n is arbitrarily close to a linear dependence on n. In particular, however fast a function $r(n)$ grows, an agenda can be defined with t_n growing as $\Theta(n \cdot r^{-1}(n))$. However, in this case it is required that b_k and $(b_k - a_k)$ grow as $\Theta(k \cdot r(k))$ i.e. the spacing constraints must be increasingly lax. Nevertheless, the result is proofed that no constant c exists such that $t_n \leq cn$ for all n, i.e. the fastest learning corresponds to a superlinear introduction time function.

Their results and methods are here considered as main reference for the algorithmic scheduling approach. In particular, the analysis presented aims at tackling two major issues only cited in the referenced work, namely the algorithm flexibility to account for possible failures of the learning procedures and the role of correlations between the units. How these two issues are dealt with is the object of the next section.

3.2 Learning schedules

The set of items to be learned are nodes in a graph, whose topology is representative for all the possible semantic connections between the items. The learning is the dynamical process of graph exploration, described by means of a *learning schedule*, namely the sequence of successive visits an hypothetical student would make to the nodes of the graph. In this scheme, the algorithm which generates the learning schedule is probabilistic, in order to be more adaptable to any learner performances or needs during the learning process, for example by taking into account failures like forgetting events. Moreover, different strategies are investigated to take into account the effect arising from the underlying topology of semantic interconnections between the items.

3.2.1 Algorithm – Time constraints

In this section it is described how the lag and spacing effects are implemented in the generation rules of the learning schedules, i.e., in the devised algorithm. The ordered sequence of nodes is defined as follows. At each time either a new node (never visited before) can enter in the sequence, or an already considered one can be repeated (subfigure (A) of Figure 3.1). In particular, at each time step, the item i to be presented to the student, i.e., appended to the learning sequence, is stochastically chosen according to three factors:

- the time, t_i, elapsed for each item i since its last presentation;
- the time, t_{new}, elapsed since the last introduction of a brand new item;
- the knowledge strength $S_i(t)$ of item i at time t, where the knowledge strength quantifies how much the i-th item is well-known. It will be defined in details in the next paragraph.

Constraints on the time window useful for reviewing an item are provided, thus taking into account the spacing and lag effects. As in a previous work [88], in order to prevent the forgetting of the item, any two successive occurrences of the same item i should take place within the given temporal interval $[a_{S_i(t)}, b_{S_i(t)}]$, whose bounds are monotonic non-decreasing function of the knowledge strength S_i. In particular, at each discrete time t:

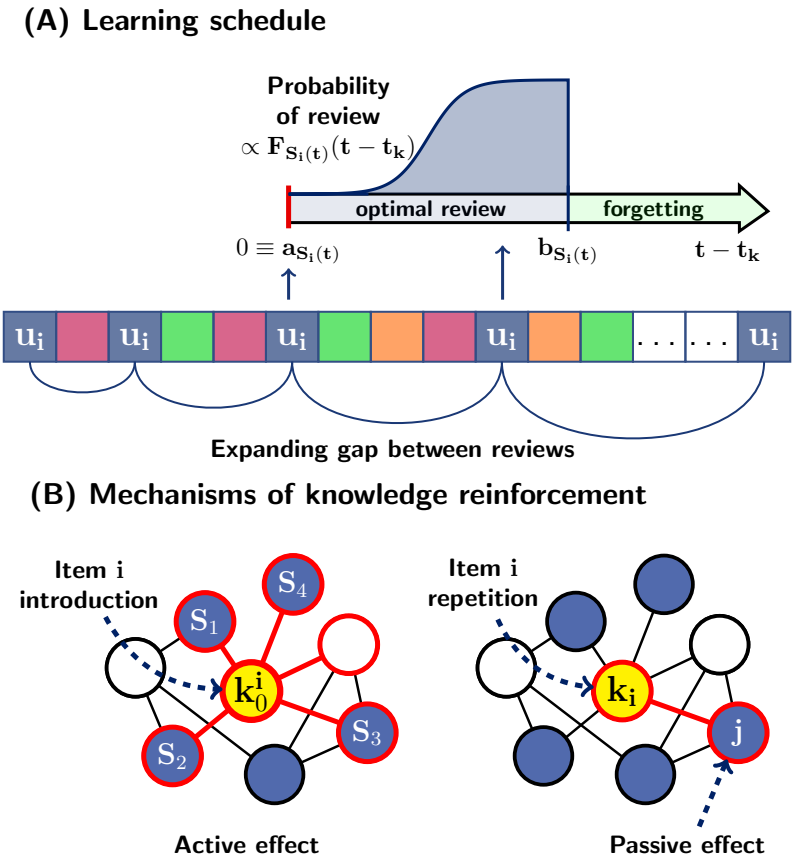


Fig. 3.1 **Model illustration.** (A) In a learning schedule the interval required between any two successive presentations of the same item i expands with the number of reviews. To this end the probability of a repetition is computed for every node already introduced as illustrated. If the k-th presentation of a node i with knowledge strength S_i occurred at time t_k, the $(k+1)$-th happens at time t with probability proportional to $F_{S_i(t)}(t-t_k)$. This function is non null in the temporal interval $[a_{S_i(t)}, b_{S_i(t)}]$, whose bounds are increasing functions of the total knowledge strength of item i. After $b_{S_i(t)}$ steps without being repeated, the item i is forgotten and has to be reintroduced. In (B) the supposed mechanisms of knowledge reinforcement are illustrated. When item i is introduced, it gains a starting knowledge value k_0^i depending on how much its neighborhood is known. This mechanism is referred to as *active effect*. Afterwards, at every successive repetition, its knowledge is reinforced by 1 and one among its introduced neighbours, say j, is randomly selected to receive a *passive* reinforcement, i.e., k_j^{pass} is incremented by a quantity α. In all the simulations here conducted $\alpha = 0.1$.

1. for each item i among the $n(t)$ already introduced in the schedule, the temporal distance since its last occurrence is evaluated: $\Delta_i t = (t - t_i)$, where t_i is the last time at which the item i entered in the sequence. If $\Delta_i t > b_{S_i}$, the item is forgotten, put into a *forgetting queue* and its knowledge strength S_i is reset

to zero. If $\Delta_i t \leq b_{S_i}$, a monotonic non-decreasing function of $\Delta_i t$, $F_{S_i}(\Delta_i t)$, determines the probability for node i to be repeated at time t.

2. the probability of introducing in the sequence a new item instead of repeating an already introduced is evaluated as $F_{new}(t) = \frac{1}{2} \cdot (t - t_{new})$.

Thus two complementary events could occur, namely an item i is chosen to be reviewed or a novel item is introduced. The corresponding normalized probabilities are:

$$\begin{aligned} P^i_{rep}(t) &= \frac{F_{S_i}(\Delta_i t)}{\sum_{u=1}^{n(t)} F_{S_u}(\Delta_u t) + F_{new}(t)} \\ P_{new}(t) &= \frac{F_{new}(t)}{\sum_{u=1}^{n(t)} F_{S_u}(\Delta_u t) + F_{new}(t)} \end{aligned} \tag{3.1}$$

In the case of a new introduction event, the oldest item stored in the forgetting queue is reintroduced, without updating t_{new}. If the forgetting queue is empty, a brand new node is introduced to the learning schedule and t_{new} is updated.

It is worth noting here that a strong approximation is made when considering what happens after b_{S_i} time steps without any new review. An abrupt forgetting is assumed, regardless of the usually supposed exponential decreasing forgetting curve, firstly introduced by Ebbinghaus [39]. The assumption made is just a first attempt to insert the possibility for some units to be missed, thus reflecting the simplification for which a unit is either remembered or forgotten, without evaluating the strength of its memory trace.

In the present scheme, many elements are free to be adapted to the learners capacities, as the functional expressions for the temporal bounds $a_{S_i(t)}, b_{S_i(t)}$ and the repetition probability function $F_{S_i}(\Delta_i t)$. In the analysis here presented, they are set as follows:

- the temporal bounds are $a_{S_i(t)} = 0$ and $b_{S_i(t)} = 2^{S_i(t)+3}$, as illustrated in Figure 3.1, subfig. (A). In so doing, it is supposed that the temporal window useful for a review to occur expands exponentially with the number of reviews [94];

- the repetition probability function is defined so that a review is more likely to happen the closer the time is to the upper bound $b_{S_i(t)}$, thus taking full advantage of the lag effect. For each item i:

$$F_{S_i}(\Delta_i t) = \frac{F^*_{S_i}(\Delta_i t) - F^*_{S_i}(0)}{F^*_{S_i}(b_{S_i}) - F^*_{S_i}(0)}, \tag{3.2}$$

where

$$F^*_{S_i}(\Delta_i t) = \frac{1}{2} \cdot \left\{ \tanh \left[\frac{LR}{b_{S_i}} \left(\Delta_i t - \frac{b_{S_i}}{2} \right) \right] + 1 \right\}. \tag{3.3}$$

In this definition, LR is the only free parameter, which stands for *learning rigidity* and fixes the function slope. In the following, it is set $LR = 2^3$.

Flexible learning

The parameter LR affects the slope of the repetition probability function, thus balancing the urgency of a repetition in the temporal width $[a_{S_i(t)}, b_{S_i(t)}]$. How its value affects the learning is here tested and reported for the case of uncorrelated items, where the knowledge strength of each item $S_i(t)$ corresponds only to the number of reviews of item i occurred up to time t. In particular, to quantify the efficiency of the process the following quantity are considered: the average number of units forgotten $\langle n_{fq}(t, LR) \rangle_t$, the average number of time steps a unit has to wait in the queue before being reintroduced $\langle t_{fq}(u, LR) \rangle_u$, and the introduction rate $n(t)$, i.e., the number of distinct units introduced as a function of the time.

In Figure 3.2, for several values of LR it is reported the corresponding repetition probability function (subfigure (a)) and the values of the above mentioned quantities. Regarding the introduction rate (subfigure (c)), the trends are fitted with a sublinear function $n(t) \propto t^{exp}$. The fitted exponents are reported in the inset. The data are contrasted with the limit case in which $F_{S_i}(t)$ is a step function of t, thus corresponding to the case $LR = \infty$. It is worth noting that an efficiency criterion based on the only introduction rate does not lead to the same evaluation of the learning performances as when the forgetting dynamics is considered. Indeed, while LR gets larger, the introduction rate monotonically increases, this corresponding to a faster learning process. Nevertheless, more units are forgotten during the procedure as well as more time is needed for them to be reintroduced from the forgetting queue. However, the

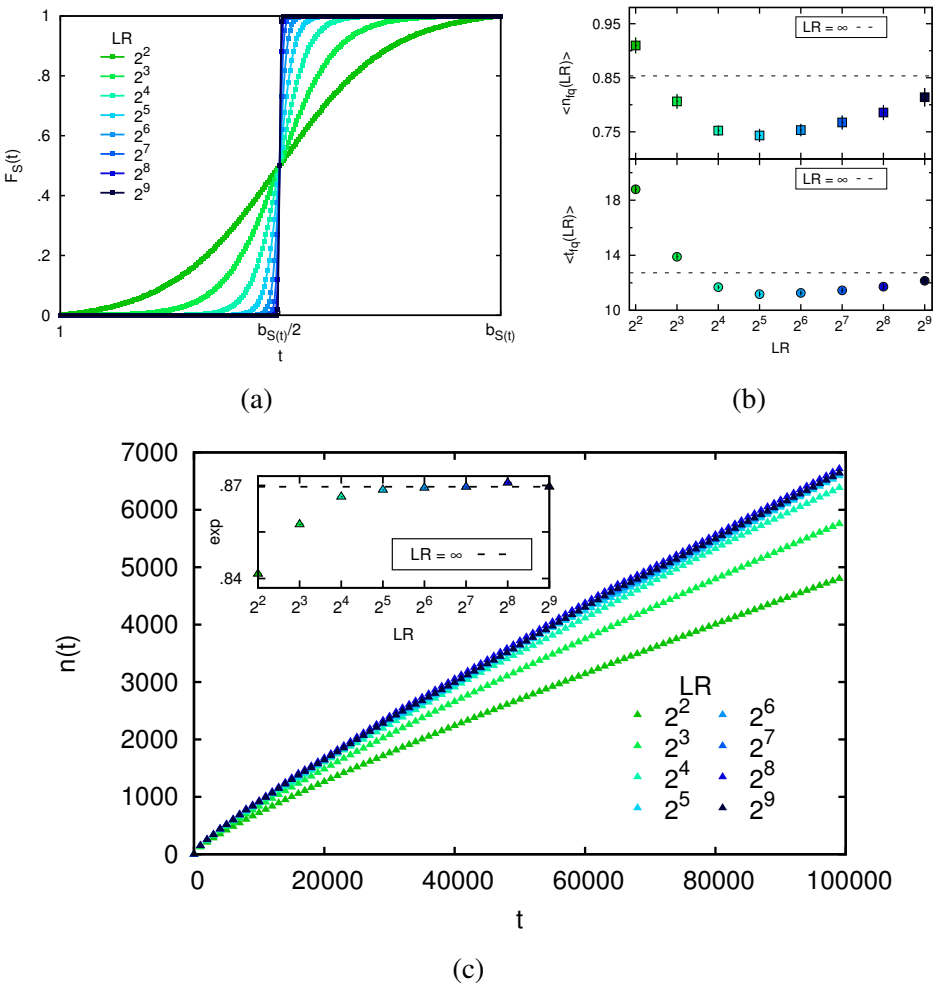


Fig. 3.2 **Learning rigidity and schedule efficiency.** In (a), different plots for the function $F_{S_i}(t)$ (Eq. 3.3) in the interval $[0, b_{S_i(t)}]$ are reported, while tuning the value of the parameter *LR*. Correspondingly, in the figures (b) and (c) some properties of the agendas obtained by running the simulations on sets of 10^4 disconnected nodes are displayed. In particular, in (b) it is reported the average number of units in the forgetting queue (subfigure at the top) and the average number of time steps a forgotten unit has to wait before being reintroduced (bottom). In (c) the introduction rates $n(t)$ are reported, with the exponent of the fitting function $n(t) \propto t^{exp}$ in the inset. All the data are averaged on 50 runs. Standard errors are reported.

functional form of the introduction dynamics (quantified by the introduction rate $n(t)$) is not affected by changing the learning rigidity parameter, remaining sublinear.

In the following analysis, the value $LR = 2^3$ is chosen, only because of the definition of the bound $b_{S_i(t)}$ and of the repetition probability function, equation 3.3.

3.2.2 Algorithm – Taking connection into accounts

The algorithm rules described so far have only taken into account the temporal constraints of the process, as they arise from the spacing and lag effects. Along with these effects, in the algorithmic scheme here proposed, the semantic connections between the units to-be-learned must be also taken under consideration.

Without entering into the details of all the possible types of connection, and their role on learning, two main mechanisms are here described and implemented, as a result of several suggestions from previous studies [32, 31, 107, 60, 59, 17], reviewed in the previous chapter. Here it is assumed that the relations between the units may indeed influence the learning rate and retention both because of a direct associative effect of the previously gained pieces of knowledge on the novel ones acquired, and because of the way the new entries are sorted, depending on their role in the whole network. These two different effects are put into the stochastic model as explained in the following subsections.

Mechanisms of knowledge reinforcement

In the previous section, the knowledge strength of item i at time t, $S_i(t)$ was introduced as a general quantifier of the extent by which the item is well-known and retained in memory. It is defined as the sum of three distinct contributions, corresponding to three different mechanisms that are supposed to lead to the acquisition and reinforcement of any item knowledge:

$$S_i(t) = k_i(t) + k_i^0 + k_i^{pass}(t) \tag{3.4}$$

where:

- $k_i(t)$: number of repetitions. It is the number of times the item i is repeated since its first introduction or since its reintroduction from the forgetting queue. Each repetition is supposed to equally contribute to the reinforce of the item in memory.
- k_0^i: *active effect*. When an item i enters in the sequence for the first time or from the forgetting queue, its starting knowledge k_0^i is a weighted average of

the knowledge acquired so far on its neighbours. The underlying assumption here is that the better the context of a new item is known, the easier learning it.

- $k_i^{pass}(t)$: *passive effect*. Every time an item j is repeated, one among its neighbours already introduced (and not forgotten), say i, is randomly selected (uniformly or with probability proportional to the weight of the connecting link, respectively in unweighted or weighted graph) and k_i^{pass} increases by a value $\alpha < 1$. This passive effect is introduced to account for the fact that the repetition of a unit may reinforce the memory of one among its already introduced nearest neighbours, as it can happen due to an explicit recall or because of a natural associative learner's behaviour.

More in details, for what concerns the active effect, it is assumed that the starting knowledge k_0^i of an item reflects the knowledge of its context, i.e., of its neighbours. It is defined by:

$$k_0^i = \max\left(1\ ,\ \text{int}\left[\langle k\rangle_{nn_i}\cdot\left(1-\frac{1}{nn_{intro}^i}\right)\right]\right) \tag{3.5}$$

where nn_{intro}^i is the number of neighbours of i already introduced and not forgotten and $\langle k\rangle_{nn_i}$ is the average no-passive knowledge strength over the set of neighbours $\mathcal{N}_i$ of item i. It is defined as:

$$\langle k\rangle_{nn_i} = \frac{1}{s_i}\sum_{j\in\mathcal{N}_i}(k_0^j + k_j)w_{ij} \tag{3.6}$$

where w_{ij} is the weight of the link connecting node i to node j ($w_{ij} = 1$ in an unweighted graph) and s_i is the strength of node i: $s_i = \sum_j w_{ij}$. With this choice, the acquisition of a low-degree unit (as a very specific word or concept) is enhanced only if the connected units have high memory strengths, while for a hub (a common word or a very general concept) the extension of the knowledge over its many linked units becomes more relevant.

Regarding the passive effect, there are of course many choices for modeling this mechanism and quantifying its range. For instance, it might contribute to k_i and be considered like a proper repetition, or it might be involved only in the computation of either the repetition probability or other units starting strength k_0^j. Its magnitude might depend on the units degrees, number of repetitions, time gaps since their last

occurrences, type of association, time distance between their first presentations into the agenda and so on. Similarly, the reinforced neighbour of the repeated node might be selected according to different criteria, e.g. proportionally to its degree, to the temporal gap since its last review, etc.

Here, the *passive effect* is introduced as follows. Every time a unit is repeated, the introduced neighbour that gains the reinforcement is randomly, uniformly selected among the available ones, if any. In this way, during the learning procedure, each node can accumulate a contribute k_{pass} exclusively from the *passive effect*, possibly reset to zero if it got forgotten. As a consequence, this contribution reveals how often the unit is recalled because of a repetition in its neighbourhood.

The particular value for the parameter k_{pass} is chosen to be the same for all the units. It is indicated with α. Tests have been done on how this parameter can affect the learning schedules proprieties, and some results will be later discussed. However, if not diversely stated, α is to be considered fixed and set to $\alpha = 0.1$.

Entry selection criteria

The order used to explore the learning environment is particularly relevant, since differently sorting the entries means differently considering the information stored in the topology. The question of finding the *optimal sorting* is indeed related to how it is usually perceived the environment in acquiring new knowledge, i.e., the way in which the information networks grow.

For instance, the early language acquisition by toddlers may underlie the formation of the observed scale-free associative networks in adults. In this case, several models for acquisition have been proposed and analyzed by studying how the normative children early semantic networks evolve. With respect to the case of language learning, throughout different word types, two main factors seem to be relevant, i.e. *the contextual diversity* and *the consistency of the context*, using the same definitions as Hills et al. have in their papers [60, 59]. The former refers to how rich the neighbourhood of a given word is, i.e., its connectivity. The latter is instead related not only to the properties of the learning environment but also to the already known words and their connections: it refers to the presumed principle that the more the neighbourhood of a word is already known, the earlier that word is acquired, thus yielding to a clustered local exploration of the network.

Starting from these results, the following different criteria are here considered to sort the entries, i.e., to select the particular brand new node to be introduced in the learning schedule:

- **Random Learning (RL)** – each new entry is randomly selected among the ones not already introduced into the agenda. This sort of selection uses no information of the network. When scale-free graphs are considered, this leads to the earlier entering of low-degree nodes.
- **Preferential Acquisition (PA)** – the new entries are chosen with probability proportional to their degrees (or strength, in case of weighted graph), thus firstly preferring the ones more *contextual diverse*. With this choice, the more frequent concepts in the learning environment (or the ones with mostly diverse context) enter earlier in the schedule.
- **Random Surfing (RS)** – inspired by the PageRank algorithm proposed by Brin and Page [23, 90]. The new entries are selected as the learner could randomly follow the connections. This could for example happen to a learner while researching about something on a encyclopedia, or on a dictionary (in this case, the connections being semantic or morphological). Formally, every time a new unit has to be chosen, with probability p a nearest neighbour of the last new one is selected among the non introduced ones, if any, with probability proportional to its degree. Otherwise, a jump is made in the network and a random node is selected with a PA step. In case of weighted graph, strengths are considered instead of degrees. The non-jumping probability p is set equal to 0.9.

The mechanism for the toddlers' early acquisitions network growth proposed by Hills et al. [59] and inspired by the *consistency of the context* principle, is here not considered. Indeed, although it could be significant in the selection of new entries, presumably leading to the earlier exploration of high clustered groups of units than that less clustered, the principle that units with more connections among the known ones are easier learned than those less related is already present in the model, through the *active effect*.

3.2.3 Quantifying the learning efficiency

On the generated sequences, two main quantities are studied to evaluate the efficiency of the corresponding learning processes. The first one is the introduction rate $n(t)$, namely the number of distinct nodes presented throughout the sequence as a function of time and not forgotten. The second variable is the graph coverage time, that is t_N such that $n(t_N) = N$, i.e., the time needed to present every node at least once and to empty the forgetting queue. For these quantities two different behaviours can be expected in the limit cases of totally disconnected and connected graphs. Because of the generation rule previously explained, and in particular the active knowledge reinforcement term, interconnections between nodes lead to a faster rate of introductions and therefore to a shorter coverage time.

However, for intermediate connectivity values, the learning efficiency does depend on both the topology of the graph explored and, for a given topology, on the criterion according to which novel nodes are to be introduced. For this, simulations on different types of synthetic graphs and on networks generated from real data are performed. In the first case, for each graph type the sequences obtained from graphs with increasing average degree are compared. For the real networks, methods of perturbation have been developed to increase and decrease the connectivity while only slightly modifying the other statistical properties, such as the degree or strength distributions. They are described in the next section.

3.3 Methods and graphs

To understand the role of the topology in affecting the learning process, i.e., in affecting the properties of the learning schedules, two main groups of graphs were considered: synthetic and real graphs.

By testing the algorithm on the synthetic graphs, it is possible to understand the role of particular statistical properties of the topologies, by tuning proper parameters in the graph generations. In particular, for different classes of synthetic graphs, their mean connectivity is tuned, while keeping fixed all the other properties.

The synthetic graph classes considered are generated through the Random graph model [41] (ER), the Barabaási-Albert model [14] (BA), the Holme-Kim [63] model

and the Uncorrelated configuration model [28]. They are all reviewed in the Appendix section A.2.

In contrast, the real graphs considered represent true semantic networks. Since they differ in generation rules, strength and weight distributions, any direct comparison of the learning schedules obtained on them would be meaningless. For this reason, to evaluate the resulting learning efficiencies, the original networks are contrasted with slightly perturbed version of them. Both the real graphs analysed and the perturbation strategies are here described.

3.3.1 Real semantic networks

With regards to the real semantic graphs, both free-associated words networks and subgraphs of Wikipedia are considered.

Free association words dataset

Among all possible semantic network representations of language, word association graphs can be considered as a proxy for how human mind stores and organizes words and related meanings [55]. Here, two datasets are used, namely the *Human Brain Cloud* [49] (HBC) dataset and *Edinburgh Associative Thesaurus* [69, 2] (EAT).

EAT dataset The *Edinburgh Associative Thesaurus* is a thesaurus of empirical word association norms [2] as they emerged through discrete associative tasks conducted in a controlled experimental environment in 1973. For the detailed data collection method, see [69]. Here, the main properties of the applied procedure are summarized.

A small set of stimulus words were defined at the beginning of the data collection procedure. The items in this nucleus set were selected according to previous studies on norms and word frequency counts. The responses obtained to these cues words were then used (with "only a minimal amount of selection" [69]) as stimuli themselves in successive experimental sessions. This sort of cycle was repeated until 8,400 stimuli were used.

Each cue word was presented to 100 different subjects. The stimuli were randomly gathered in groups of 100 words for each participant, who, as a consequence,

contributed with 100 responses. Similarly, due to this data collection procedure, each word used as cue resulted to have an upper-bounded out-going strength of 100.

The data of word associations are free available in [2]. From them, a undirected weighted network is generated, without self-loops. The final graph is composed by 23,219 nodes and 289,116 associations. The experimental set-up constraint about the number of possible responses for each cue results in a peaked strength distribution, as shown in Figure 3.3, subfigure (B)).

HBC dataset The *Human Brain Cloud* database is the largest word association database available at the present [55]. The associations have been collected through a multi-player web-based game [49], designed without any particular scientific purpose of analysis. In this game, players are asked to response with a target word to a cue word randomly proposed by the system. No control is performed on the number of participant, nor on the number of associations with which each player contributes. On the words proposed as cues, a sort of user-based filter procedure is designed, so that only the words with a minimum *quality level*[1] are used as stimulus. The game indeed started with only a word, while the internal dictionary from which the stimulus words are extracted automatically grows by gathering the answered responses at the end of game sessions.

After a data collection within a period of a year, the data set consists of around 600,000 words and 7,000,000 associations. Of these words, the ones have been discarded which have been reported by the same users as *inconsistent*, e.g. because misspelled or offensive. Further filtering of the data by removing the non valid words selected with the same criterion as the cues were was performed by Gravino et al. [55]. In the cited work, the authors also proved the robustness and reliability of the dataset, by recovering significant correspondences with other free word association graphs, based on controlled linguistic experiments.

The final network, by courtesy provided by Gravino [55], consists of around 90,000 words and 6,000,000 associations. In this work, while edge weights are preserved, any information about link directionality is lost, thus treating the graph as undirected.

[1] They are the same participants who are asked to attribute to each word a validity score, based on its popularity.

Wikipedia subsections

Of the entire Wikipedia graph [7], some subgraphs corresponding to the following particular scientific area are extracted: Physics, Mathematics and Chemistry. To this end, the MediaWiki API [4] were used[2] as described. First, the list of thematic Wikipedia article titles was fetched by enquiring the API for the corresponding scientific area, i.e., by restricting to the corresponding category, e.g., *Category:Physics articles by importance*. Then, each page referring to the titles collected was scanned for the included links to other pages. Pages containing talks, templates and categories were not taken into account as well as connections toward pages not belonging to the subsection.

In the following, the particular case of the Physics subgraph is discussed, as representative of all the Wikipedia subsections considered. Indeed, while slightly different for size, the three above mentioned subgraphs are statistically similar and similar results have been obtained simulating the learning agendas on them. The Physics subsection is chosen as representative only because it is the biggest among the three. All the results obtained on the other sections can be however found in the Supplementary Information file to [96].

The Physics subgraph is treated here as undirected, and results as composed by 16426 nodes and around 424k edges. From this graph also some inner k-cores (see Appendix A) are extracted to evaluate the role of the least connected node in affecting the process.

Perturbation of real-world graphs

The real graphs considered are diverse for network order, topological properties and type of data represented. Thus, directly contrasting the results obtained by running the learning algorithm on them could is not meaningful. Instead, some techniques are devised to slightly modify the original topologies, thus obtaining some perturbed versions for each of the starting graphs. By contrasting the learning efficiency gained on these perturbed topologies with what obtained from the original graph, meaningful insights could be acquired.

[2] Date of access: 22/11/2013

The perturbation procedure is here described. Starting from a real-data based graph, a predefined percentage of links is created or deleted according to the following criteria. When it is required to remove some connections, they are randomly selected and deleted, regardless of their weights or of the degrees of the connected nodes. As a main consequence, some disconnected components might emerge.

In adding links, two different strategies are implemented. In a first case, two reciprocally disconnected nodes are randomly selected and a connection is created between them, regardless of their distance on the graph. As a consequence of this rule, the transitivity of the graph is strongly affected, rapidly decreasing as new edges are inserted. According to a second procedure, a node is randomly selected and a new connection is created with one among its second-neighbours, in this way reducing the effect on the graph transitivity.

Both in case of removal or addition, the new link weight is possibly assigned by sampling the original weight distribution. In particular, an edge in the original network is randomly selected, and its same weight is assigned to the new link.

Figure 3.3 displays the original strength and degree distribution for all the real graph considered, together with the same distribution after perturbing the original graphs.

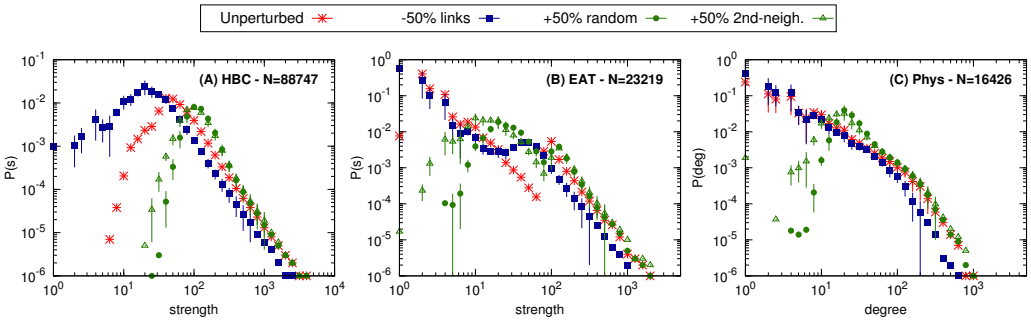


Fig. 3.3 **Strength and degree distribution for original and perturbed real graphs.** data are displayed which refer to (A) HBC graph, (B) EAT graph and (C) Wikipedia Physics subsection graph. Red symbols refer to the unperturbed graphs. Starting from each unperturbed structure, 50% of links were randomly removed (blue squares), randomly added (filled green circles) or randomly added only between second neighbours (empty green triangles). The data are averaged over 10 different realizations of each perturbation procedure. Standard deviations are shown.

All the original topologies display degree/strength distributions with large tails, corresponding to the presence of hubs in the networks. When perturbed, the nodes

least connected are largely involved, with the appearing of peaks in the low-middle connectivity range as soon as novel links are added.

3.4 Results

3.4.1 Learning on artificial topologies

From the analysis of the learning schedules obtained by applying the algorithm on synthetic networks, a global insight into the role of the graph topology and its connectivity properties can be gained, especially by investigating the coverage time.

In Figure 3.4 the average coverage times is shown as obtained for the following synthetic network types: random graphs [42] (subfigure (A)), scale-free BA [14] graphs (subfigures (B)) and graphs generated with the Uncorrelated Configuration Model (UCM) [28] (subfigures (C)-(D)). For each graph type, the results obtained for the three entry selection criteria earlier defined are displayed with different colors.

Firstly, it can be observed that a scale-free topology together with no random criteria of exploration leads to optimal learning performances, i.e., the fastest, for intermediate average connectivities. The improvement in the coverage time is even more meaningful in graphs with the same maximum degree but a larger fraction of hubs, as it emerges by comparing the UCM networks with two different exponents of the degree probability distribution, reported in subfigures (C) and (D). With regard to the selection criteria, an efficiency gain is achieved in the scale-free graphs when they are locally explored, namely when the random surfing criterion is used.

For the particular cases of random ER graphs and scale free graphs based on BA model, it has been investigated the dependence of the coverage time on the size of the systems, i.e., the networks orders N. More in details, it was studied the variation of the rescaled coverage time (t_N/N) as a function of N for some fixed values of the average connectivity ($\langle deg \rangle$). For both the graph types the following trend is found:

$$\frac{t_N}{N} \sim C(\langle deg \rangle) \log N \tag{3.7}$$

that is, the rescaled coverage time scales logarithmically with the network order. The coefficient is a function of the mean graph connectivity. This functional dependence is very slight for ER graph, while more significant in the BA. In these graphs

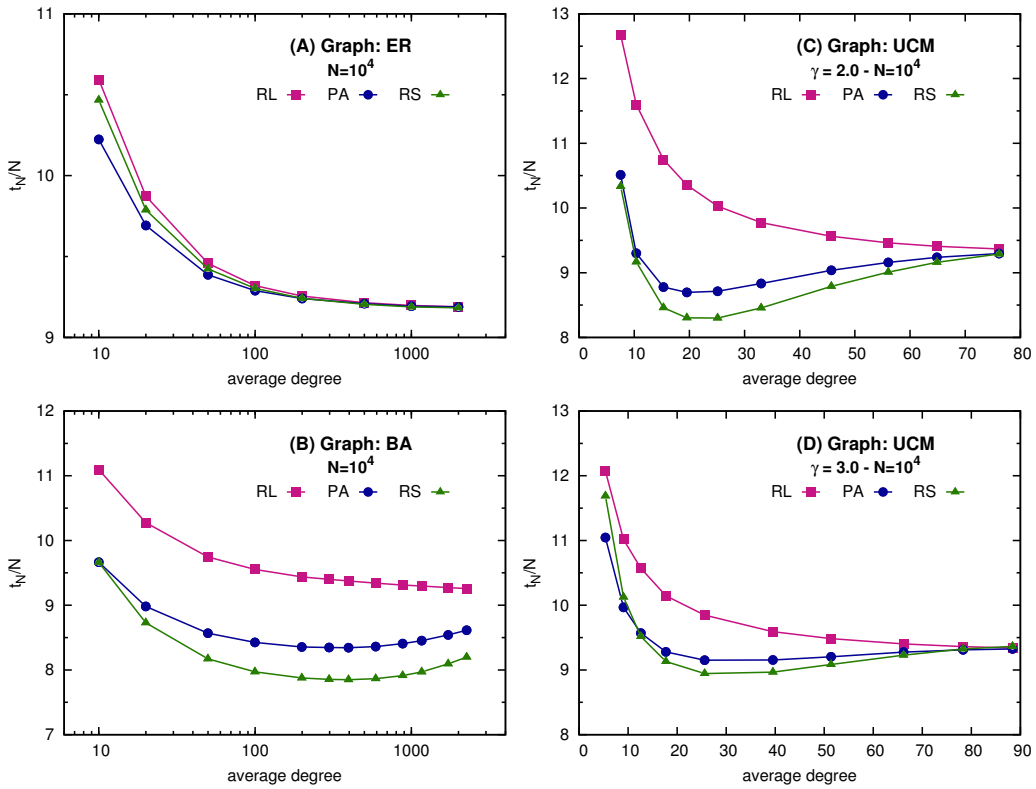


Fig. 3.4 **Coverage times on synthetic graphs.** In the figures, the mean coverage times scaled to the network order $N = 10^4$ as a function of the network average degrees are shown for different synthetic graphs, generated according to the (A) Erdös-Rényi model [42], (B) Barabási-Albert model [14], (C) uncorrelated configuration model [28] where $P(deg) \propto deg^{-\gamma}$ and $\gamma = 2$, (D) same as in (C) with $\gamma = 3$. The data are averaged over 10 different graph realizations and 5 learning agendas for each of them. Standard errors are also reported but they are covered by symbols. Different colors refer to the three criteria used to select the entries: random learning (RL, magenta), preferential acquisition (PA, blue) and random surfing (RS, green). Note the logarithmic scale of the horizontal axis for (A) and (B).

$C(\langle deg \rangle)$ is a decreasing function of the mean connectivity. It turns out that, while in ER the coverage time increases with the system size, in the scale free graph the learning process is differently affected by a system size modification, depending on the average link density.

It is worth noting here that it could appear as barely meaningful testing the algorithm on scale-free graphs with very high values of connectivities, as for BA networks with a high average degrees with respect to the total size of the network, because of the poor heterogeneity in the node connectivities. Nevertheless, some

insights could still be learnt. First, it is exactly by compressing the degree distribution that the need for a structured, heterogeneous topology appears in order to improve the learning performance. In addition, despite the narrow distribution of degrees, it can be tested how much different exploration criteria could still lead to significantly different outcomes. Finally, it is always possible to contrast the results obtained on BA networks with very high average degree with equivalently connected ER graphs. Indeed, from this contrast it is obtained that even a slight heterogeneity in the degree distribution leads to an improved learning performance when compared to completely homogeneous networks.

In order to investigate the role of transitivity in determining the learning efficiency, simulations are conducted on graphs generated according to the model proposed by Holme and Kim [63] (see Appendix A). Indeed, this model allows to control the average clustering coefficient by properly tuning a parameter in the graph generation, while preserving the scale-free degree distribution. Results for the coverage times on graph with different minimum degree are reported in Fig. 3.5, for both the PA and RS entry selection criteria. Changes in the transitivity do not affect the learning

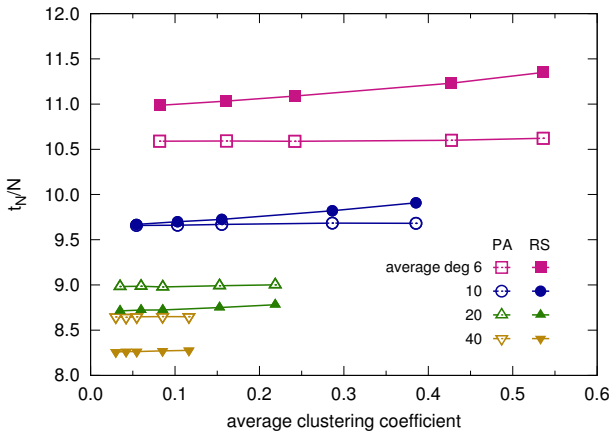


Fig. 3.5 **Coverage times and graph transitivity.** The figure shows the average coverage times scaled to the network order ($N = 10^4$) obtained on scale-free graphs generated according to the model proposed by Holme and Kim [63] as functions of the average clustering coefficient. Different colors refer to graphs with different minimum (and thus also average) degree, while empty or filled symbols distinguish data obtained when respectively the PA or the RS entry selection criterion is used.

procedure when PA criterion is used, while a higher clustering coefficient in the network hinders the learning procedure if the RS rule is implemented. This result

could be explained by the fact that, when RS criterion is used, the learner tries to explore locally the graph, as far as any new unexplored nodes are available. If the graph is largely clustered, clusters are quickly covered and the learner is forced to jump somewhere else in the graph, thus loosing possible reinforcement effects from the context.

Along with the study of the coverage time, another insight into the dynamics of the learning schedule construction process is given by looking at the introduction rate $n(t)$. In Figure 3.6 (subfigure (A)) results on random and BA graphs with similar average degree are compared. In subfigure (B) the data refer to graphs generated with the UCM model with low, intermediate and high values of average connectivity, and exponent $\gamma = 2$ in the power law degree distribution $P(deg) \propto deg^{-\gamma}$.

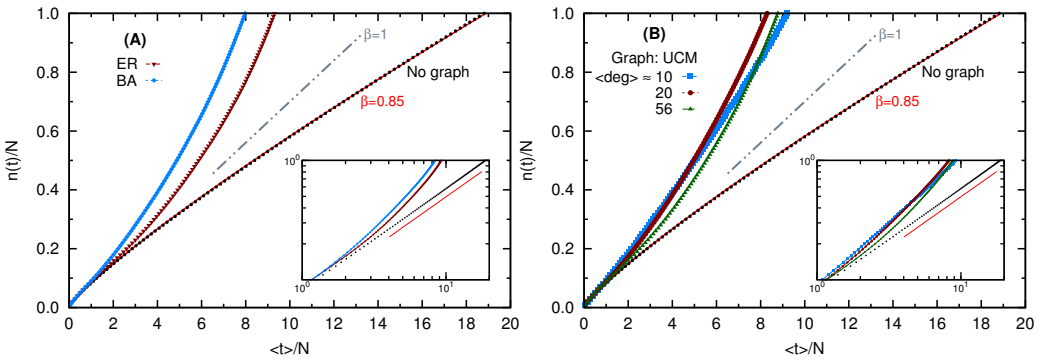


Fig. 3.6 **Introduction rate on synthetic graphs.** Fraction of distinct nodes introduced as a function of the (rescaled) average time needed to cover them. In figure (A) we report data obtained on Erdös-Rényi model [42] and Barabási-Albert model [14] graphs with average connectivity $\langle deg \rangle \sim 100$. In (B) the graphs considered are generated according to the uncorrelated configuration model [28] with $P(deg) \propto deg^{-\gamma}$ and $\gamma = 2$ and different average connectivities. In all the cases, the graph size is $N = 10^4$ and the algorithm used to select the entries is the random surfing RS. In both (A) and (B), with black dots we report the introduction rate for an equivalent set of uncorrelated items. It is fitted with a sub-linear curve $y \propto x^{\beta}$, with $\beta = 0.85$. The fitting curve is shown with red line in the main graph, while in the insets we report an eye-guide power-law with same exponent. An eye-guide linear ($\beta = 1$, grey line) curve is also reported in the main figures. The insets axes are in log-log scale. The data are averaged over 50 agenda simulations. Standard errors are reported, though not visible at the plot scale.

In both figures, the data are contrasted with the results obtained on an equivalent set of completely disconnected nodes (and a linear trend is also reported for comparison). For uncorrelated items, the introduction rate turns out to be a sub-linear function

of time ($n(t) \simeq t^{\beta}$, with $\beta < 1$), in accordance with Heaps' law [58]. Instead, for items embedded in a graph, two different behaviours can be identified. As long as the graph is largely unexplored, the introduction rate has the same trend as in the case of disconnected items, namely sub-linear. Later on along the learning dynamics, new items are introduced with higher frequency, featuring a super-linear tail for the introduction rate, i.e., $n(t) = c^* t^{\gamma}$, with $\gamma > 1$ and $c^* \ll 1$. It is worth underlining that, for a short time interval, such a super-linear rate is still compatible with the schedule constraint that at most one brand new unit can be introduced at each discrete time. The origin of this super-linear behaviour is related to the active effect contributing to the knowledge strength of each item. Indeed, when a significant fraction of items have already been introduced, new items typically enter the schedule with higher and higher knowledge strengths, thus requesting longer intervals before they need to be reviewed, allowing in this way the introduction of further new items.

Active and passive effects on scheduling

Further tests have been conducted to isolate and validate the contributions of the active and the passive effects on the learning efficiency of the generated schedules.

In order to better analyse the role of the active effect on the learning efficiency, schedules have been generated in which each node enters the agenda with a pre-assigned k_0^i value, in this way keeping fixed and independent of the dynamics the total knowledge reinforcement gained throughout the procedure. In particular, for both the cases without passive effect and with a passive contribution ($\alpha = 0.1$):

- a learning schedule on a optimal UCM [28] graph is generated (where the $deg_{MIN} = 7$ and $\gamma = 2$ in $P(deg) \propto deg^{-\gamma}$), using the RS entry selection criterion;
- the k_0^i with which each node was introduced in the previous agenda are reshuffled on the entire graph. Then a new schedule on the same graph is simulated, without recomputing the active effect, rather using the pre-assigned values. The resulting agenda is referred to with *UCM resh.*;
- as in the previous step, a further schedule is simulated starting from the same preassigned k_0^i but now considering as underlying an ER [42] graph,

with average degree around 500 (this value corresponding to a random graph yielding to the best performance, as reported in § 3.4).

In this way, while fixing the absolute values of the knowledge strength entering throughout the schedule because of the active effect, any correlation of the individual values k_0^i from the underlying graph is destroyed.

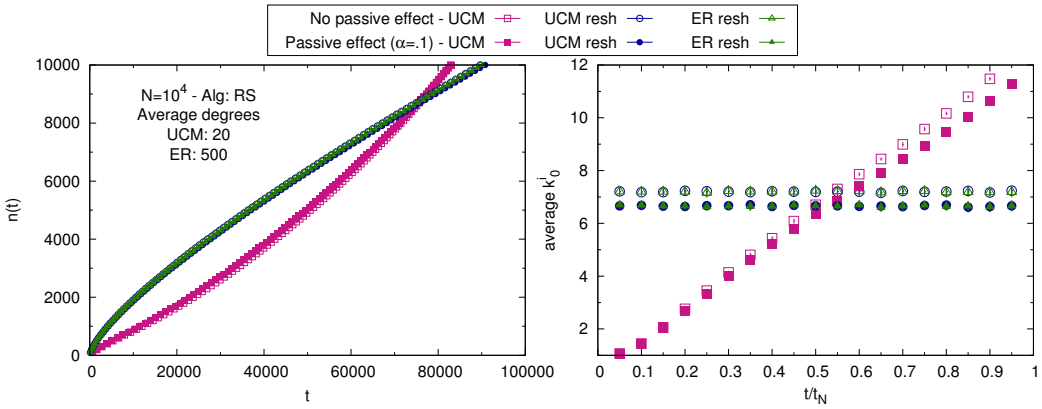


Fig. 3.7 **Results of the $\mathbf{k_0^i}$ reshuffling procedure.** Two series of 50 learning agendas are generated on a UCM graph with average degree 20 and $\gamma = 2$ (UCM, pink) without passive effect (empty symbols) and with it, setting $\alpha = 0.1$ (filled symbols). At the end of each schedule creation, the values of k_0^i are reshuffled over the nodes. Then, new agendas are generated both considering the same underlying UCM graph (UCM resh, blue) and a ER graph (ER resh, green) with average degree 500. In both cases, the nodes enter the agenda with the preassigned k_0^i-s. Of the resulting data, it is reported the introduction rate $n(t)$ (on the left) and the average k_0^i of introduced nodes as a function of their introduction time (subfigure on the right). In all the cases, the RS criterion is used to select the new entries.

The results are reported in Fig. 3.7. Since, on average, in the reshuffled cases the nodes enter the agenda with a higher starting knowledge strength, the introduction rates is initially faster in these cases. However, the final coverage times are higher than in the schedules without reshuffling. Moreover, in the reshuffled cases, no meaningful difference appears if the underlying topology is a random graph rather than a scale-free one. This means that all the scale-free graph properties useful in enhancing the learning procedure are ineffective if the active effect is separated from the dynamics.

Regarding the passive effect, simulations have been done on UCM graphs, with different proportion of hubs, namely with degree distribution $P(deg) \propto deg^{-\gamma}$, and $\gamma = 2.0, 2.5, 3.0$. In particular, for the usual fixed order $N = 10^4$ and for the only

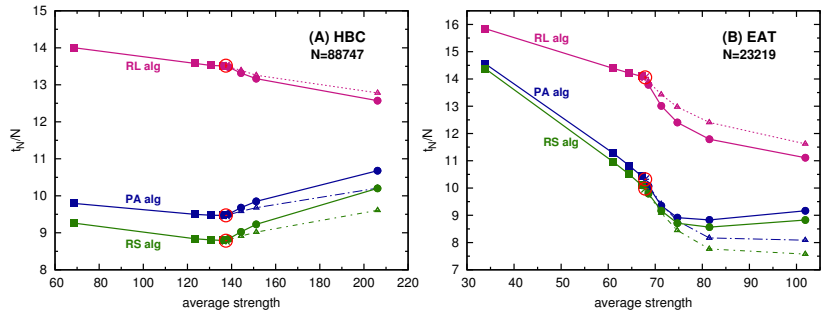
entry selection criteria PA and RS, the agendas have been simulated with a reduced passive effect, $\alpha = 0.05$, and without it, $\alpha = 0$. Two main results are obtained. First, a reduction in the passive effect does not affect the coverage time uniformly over the range of connectivities. When different values of α are used, the spread in the resulting coverage times is larger for highly connected graphs, regardless of the exponent γ of the degree distribution. This can be explained as follows. With high connectivity, the passive effect affects in a quite uniform way the entire network, i.e. the higher α, the longer the delay in the need of repetitions, and thus the minor the coverage time. A second observation regards the results in case of low average degree. In this cases, the degree distribution plays a role in determining the learning efficiency. In fact, while a reduction of the passive effect parameter α still leads to a (slightly) increase in the coverage time for $\gamma = 3.0$, the same reduction enhances the learning when $\gamma = 2.0$. A no-null passive effect, delaying the repetitions, could indeed lead to a reduction in the k_0^i of the new entries, if they are low-degree nodes. This can explain the differences between different exponents, since for the same average degree, the smallest the exponent γ, the smallest the allowed minimal degree in the network.

3.4.2 Learning on real semantic networks

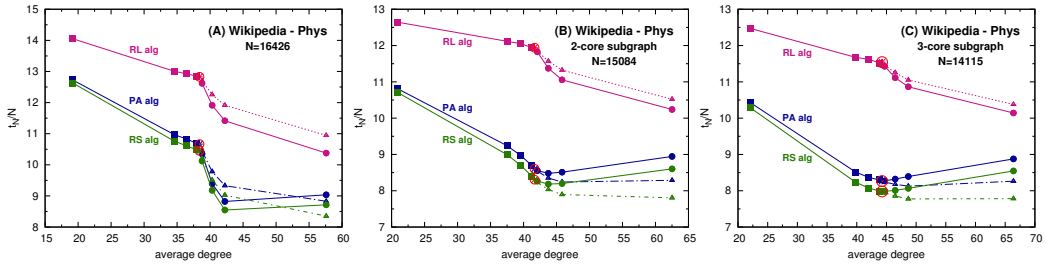
The coverage times resulting from simulations on real-world graphs and their perturbed versions are shown in Figure 3.8. In the subfigures at the top, the coverages times obtained on the two free-associations words graphs generated from the Human Brain Cloud [49] and the Edinburgh Associative Thesaurus [2, 69] are reported. At the bottom, data refer instead to the subgraph in Wikipedia corresponding to the Physics subsection and some of its first inner cores.

As for the synthetic graphs, the random learning algorithm for choosing the new entries does not lead to meaningful performances, the coverage time monotonically decreasing as the connectivity enlarges. On the contrary, when the information stored in the topology is used to more shrewdly select the novel nodes, the minimal coverage time is achieved for intermediate connectivities.

More interestingly, the structures leading to the optimal performance coincide with the original HBC graph (subfigure (aA)) and with the original Physics graph,



(a) **Free-associations graphs:** Weighted networks generated from (A) HBC[49] and (B) EAT[2] datasets.



(b) **Wikipedia Physics subgraphs:** (A) entire section and its (B) 2-core and (C) 3-core subgraphs.

Fig. 3.8 **Coverage times on real-world graphs.** It is reported the coverage times obtained by simulating the learning agendas on some real-world graphs (red circled points) and on three perturbed versions of them. The real-world graphs are indicated in the subcaptions corresponding to the two subfigures. In each figure, the squares refer to data resulting on graphs with reduced connectivity, obtained by randomly selecting and deleting different amounts of links in the original graphs. With circles and triangles data are reported when two different procedures for increasing the connectivity are considered. In the first case (circles, solid line), links are created by randomly selecting pairs of unconnected nodes. In the latter (triangles, dashed line), new links are added only between second-neighbor nodes. In all the cases, the fraction of links deleted/created are equal to 0.01, 0.05, 0.1 and 0.5. Different colors refer to the three criteria used to select the entries: random learning (RL, magenta), preferential acquisition (PA, blue) and random surfing (RS, green). The data referring to the unperturbed graphs are averaged over 50 agendas. In all the other cases, for each type of perturbation procedure and percentage of edges added or removed, 10 different perturbed versions of the graphs are generated and 5 agendas are simulated on each of them. Then, the averages are done over the 50 aggregated agendas realizations. Standard errors are reported, though not visible at the plot scale.

when the least connected nodes are removed, i.e., when the inner cores are considered and treated as unperturbed new graphs.

Based on the results obtained, two main observations can be made. First, it emerges a clear difference between HBC and EAT cases. The great difference in the strength distribution between the two graphs can help explain what found. Indeed, as shown in Figure 3.3, before any perturbations, EAT graph has a fare larger fraction of poorly connected nodes with respect to HBC. Only if the graph is positively perturbed, i.e., further links are added, its topology becomes more efficient for the learning algorithm. Interestingly, exactly the same result are obtained when the graphs are considered as unweighted.

The investigation of the results obtained in the Physics subgraph further confirm the assumption that the presence of too isolated nodes is key in hindering the learning efficiency. Indeed, as soon as the leaves are removed from the original graph, the topology of the Wiki subgraph becomes closer to the optimal one, with respect to a further increase of the number of connections, as can be seen by comparing subfigures 3.8(b).

It is worth noting that the improvement in the learning efficiency as soon as the internal cores are considered is not only an issue of accessibility, as it is for random walk type processes. Indeed, in the present case, since direct access to a node is allowed when needed for the information retention, the main drawback for the least connected node its their poor context, which makes their retention in memory less robust, thus requiring more repetitions.

This finding can be used in future to suggest a topological reorganization of Wikipedia subgraphs resulting in an optimization of thematic learning paths. Similar results, although not reported here, are obtained when other subsections of the Wikipedia graphs are considered, such as Mathematics and Chemistry subsections. Coverage times obtained on these subgraphs are exhaustively reported in the Supplementary Information document of [96].

Finally, by looking at the data acquired when the two positive perturbation procedures are implemented (circles vs triangles points in the figure), it can be concluded that it is not the average connectivity that triggers the most efficient learning performance, rather the relative presence of poorly connected nodes with respect to the hubs.

3.5 Discussion

The analysis presented in this chapter has focused on the role of the topology of complex information and knowledge networks when generating efficient learning schedules for the items they embed. To this end, a general class of stochastic algorithms is proposed to sequence the introductions of the different items and their reviews over time, while satisfying some constraints on the best timing, as they can be derived from previous results of cognitive science research. Furthermore, it is studied how the topological structure representing the complex semantic between the items to be learned can affect the learning procedure. In particular, it has been investigated how different statistical properties and topologies of the graphs in which the items are embedded affect the process, as well as the ways such graphs should be explored while introducing new material in order to achieve efficient learning paths.

The main result obtained is that some topologies lead to optimal learning schedules, i.e., schedules that minimize the learning time while preventing forgetting episodes. They are small-world, scale-free structures, in which the relative number of hubs and low-connected nodes are balanced. In fact, structures with either too many hubs or poorly connected nodes hinder the learning process. In the first case, the context for items is indeed too large to take advantage of it. In the latter case, the more specific and low connected the nodes, the more difficult it is to access them or to achieve a gain in the knowledge reinforcement throughout the learning process. Furthermore, the order through which the networks are explored as new items are introduced in the agenda is essential for taking full advantage of the topology features, a random exploration turning out to be ineffective in eliciting the information stored in the graph.

Finally, a very interesting outcome of this study is that the real-world graphs considered here, the Human Brain Cloud word-association network and the Wikipedia subgraphs, turned out to be almost optimal with respect to some perturbations of their topological structures. Other graphs not emerging from user free activities, as the graph build on the Edinburgh Associative Thesaurus, do not show *optimal* topological characteristics. This points to a subtle link between the way in which humans organise their knowledge, i.e., the structure of the knowledge space, and the way in which the information could be retrieved, for instance through a learning path.

However, many aspects have not been taken into account. Indeed, to drive the exploration of the spaces only the topological properties of the nodes embedded in the knowledge spaces has been considered. Nothing is assumed on how the *content* of each knowledge unit could influence the process. In fact, the order of exploration might constraint the section of the knowledge space actually available for the learner, who could need to firstly cover preparatory materials to have the possibility to explore new broad region of the knowledge space. Moreover, it could be needed to differentiate among the paths, some possibly being more *educationally preferable* than others. Different mechanisms would allow to implement such situations into the algorithm proposed, for example by adding attributes of directionality to the edges or by limiting the section of the space explorable at each step. Still, a greater insights must be firstly gained into the real behaviour of learners in true knowledge spaces. This is exactly the aim of the work presented in the next chapters.

Chapter 4

Free exploration of knowledge spaces

In the previous chapter a class of algorithms has been presented which allow to explore a knowledge space while respecting some temporal constraints on the most suitable times for reviewing old material or introducing novelties. Moreover, some assumptions have been made on how the learning process could be hindered or enhanced by possible semantic relationships between the item to-be-learned, depicted as nodes in a complex network. With all this, it has been possible to investigate the performance of different topologies, isolating the key factors in enhancing the process.

However, beyond the topology, the navigation behaviour of the users on the Web is not strictly a random exploration of the space if other information are taken into account, as for the *information content* of the nodes visited. Indeed, many results [111, 102] have already pointed out that patterns and strategies can be found in different information-seeking tasks of Web and Wikipedia users which do not mirror memoryless dynamics, as already reviewed in section 2.2.

In this chapter, a work is presented which moves in this proper direction and recently published [97]. It is addressed whether regularities are observed in the way users surf Wikipedia, by tapping on the recent release of data about Wikipedia readers, namely the Wikipedia Clickstream [115]. These datasets, covering some months during 2015 and 2016, provide large sets of (referer, resource) pairs extracted from the request logs of Wikipedia on a monthly bases. There, "*a referer is an HTTP header field that identifies the address of the webpage that linked to the*

resource being requested" [115], while a resource can be any of the pages in the main namespace of the Wikipedia, i.e., all the articles pages and the MainPage.

The analysis prosed in this chapter is based on the particular release of the dataset regarding the English Wikipedia Cliskstream (EWC) gathered during February 2015. From these data, the graph of the actual traffic flows streaming in and within Wikipedia has been derived. Moreover, since no direct data about the navigation histories of individual users are yet available, the EWC dataset allows for the construction of first-order Markov chains, where the transition probabilities are given by the traffic flows. These simulated paths could be considered statistically legitimate proxies for the real paths navigated by the users [102]. Both the dataset description and the rules devised to generate the simulated users' paths are presented in the next section.

It follows the description of the procedure devised to map the Wikipedia pages into a more abstract space, where to look for meaningful patterns of the users. Indeed, the analysis presented here aims at characterizing the emerging paths from a semantic point of view. To this end, a vectorial representation is introduced in which each Wikipedia page is represented by a vector of features in a 13-dimensional space [70, 102] whose dimensions correspond to broad Wikipedia topics/main categories [6, 3]. These very general subjects are here treated as coordinates of a topical abstract space in which the simulated users' paths are studied. The vector coordinates for each page in this space are computed so that the weights are proportional to the semantic relatedness of each topic with the page's parent categories.

Finally the results are presented of the analysis of users' path, based on the semantic mapping. To gain a deeper insights into the results, the data are also contrasted with paths originated in a very different task, like the goal-oriented walks of Wikispeedia [111] players.

4.1 Navigation paths on Wikipedia

4.1.1 The English Wikipedia Clickstream dataset

The main dataset here considered is the English Wikipedia Clickstream [115], released on February 2015. The dataset includes 22 millions of aggregated requests

of articles in the main namespace of the English Wikipedia, together with their *referers*, i.e., the webpages from which the requests were performed by the users during the month of February 2015, and the number of occurrences of each pairs, if exceeding 10 requests. More in details, the data were extracted from the request logs of Wikipedia and aggregated so that the referer can correspond to an article or to an external source. In the first case, the article title is given, whereas if the page is not an article, it is explicitly reported the proper category among the following: *google*, *twitter*, *bing*, *yahoo*, *facebook*, *wikipedia* (any page in Wikipedia different from an article), *internal* (any page belonging to a different internal Wikimedia project), an *empty* referer, or *other* for any different referer.

The available data in the dataset are then not raw, but already aggregated. In their last and available form, only the requests to the Wikipedia server for the pages in the main namespace are recorded and already filtered (for example by removing pages with less than 10 requests from clients clients who made too many requests). Thus, based on the information provided [115] with the dataset, only the actual clicks produced by the users are included.

Recently, several works have been proposed based on the EWC datasets. The transitions counts were used mainly to investigate how the position of links in the articles can bias the users' browsing behaviours [72, 73], to validate different link recommendation algorithms [98] and also to extract semantic relatedness between words [87, 35].

Here, the article to article counts are instead the basis for simulating real users unconstrained navigation paths on Wikipedia and thus identifying possible recurrent meaningful patterns on a semantic, abstract level of description.

To this end, the original dataset has been cleaned as here described. The requests to non existing articles (e.g., articles requested by following any *redlink*[1]) were removed. All the other (referer, resource) pairs were kept, regardless of the fact that an hyperlink exists directly connecting the two articles. The MainPage of Wikipedia was treated as an external source of navigation, even if it appears i the original dataset in the main namespace in the Wikipedia system as the other articles. As a consequence, it was added to the set of external sources previously listed (in the following: *mainpage*). This choice was due to the fact the the MainPage is often used

[1]It is worth reminding that a red link represents a link to a page that is either non-existent or deleted.

as a starting point for a research or navigation in the online encyclopedia similarly to any external web search engine, and it not semantically representable as done for the other articles (and as it is described in the next paragraph).

After cleaning, the dataset includes 3,087,211 articles. The requested {article, article} pairs are 14,076,289. Among these, the pairs of pages connected by a hyperlink are around 88%. The requested {external-source, article} pairs are 8,231,312. An illustration of the different fluxes coming from the external sources is reported in Fig. 4.1(A).

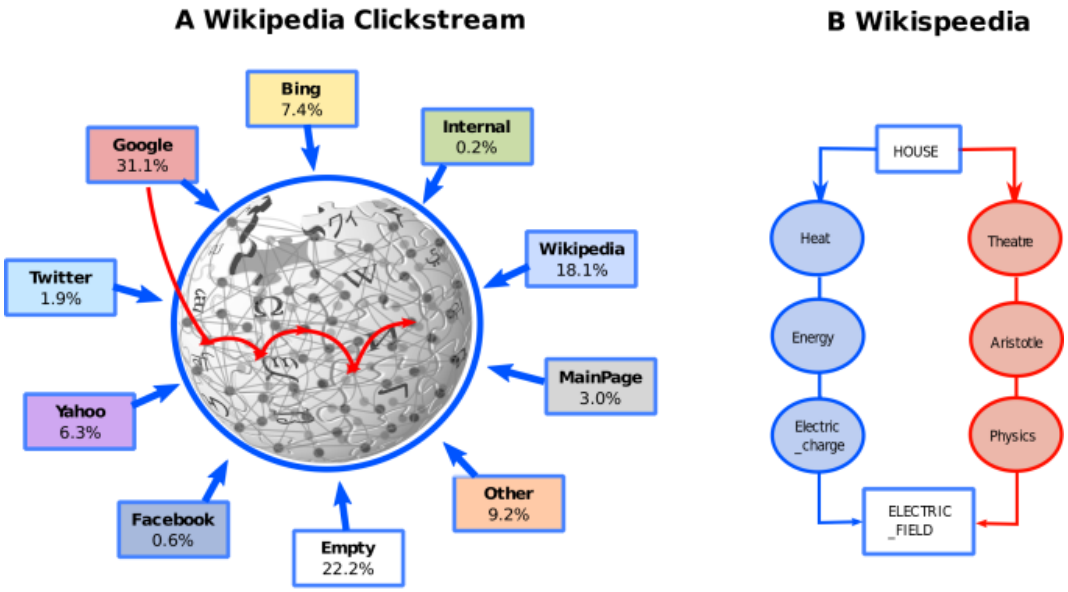


Fig. 4.1 **Datasets under consideration.** In (A) the English Wikipedia Clickstream dataset is illustrated. The 9 different external sources plus the MainPage are displayed with the fraction of flux outgoing from them. The paths considered in the present analysis start from one of the 9 sources to randomly walking over the Wikipedia articles accordingly to the transition counts provided by the dataset. As an illustration, in red line a decorative path starting from Google and them jumping on different pages of the Wikipedia graph is reported. (B) Two examples of paths followed by players of the Wikispeedia game, whose task was that of navigating on a reduced version of Wikipedia from a given starting page to a given target one (from *House* to *Electric_Field* in the example).

4.1.2 From EWC to random walks on Wikipedia

The EWC dataset provides the transition counts between pairs of Wikipedia pages. This information can be used to drive a random walker, whose paths over the

Wikipedia pages are here considered as proxies for the real behaviours of users who navigate Wikipedia. In doing so, it is assumed that the navigation over the encyclopedia can be represented by a stochastic process without memory. Indeed, while it was proven [102] that the human navigation processes are better modelled by second (or third in free navigation) order Markov chains in topical and abstract level of description, still memoryless models are statistically legitimate to simulate human navigation on a page level, as done in the present analysis.

To take full advantage of the available dataset, the starting pages of the simulated walks are selected based on the traffic flowing into Wikipedia from the external sources. On the other hand, every page can be the last one of the walk with a probability proportional to the net difference between the incoming (from other pages and from external sources) and outgoing (only towards other wikipedia articles) traffic flux on that page.

More in details, the random walks on Wikipedia based on the EWC dataset are simulated as follows. Each of the different external sources listed in the previous paragraph is in turn selected as the external origin of the simulated walker. This is the starting point of the walk.

- The first node after the origin is randomly selected among the ones reached by the source, with a probability proportional to the incoming flux;
- on every node encountered across the walk two complementary events can occur: either the walker takes one more step or the walk stops. The stopping probability at node i is defined as:

$$prob^{i}_{stop} = \max\left(0, 1 - \frac{s^{int}_{out}(i)}{s^{int}_{in}(i) + s^{ext}_{in}(i)}\right) \tag{4.1}$$

  In the above definitions, $s^{ext}_{in}(i)$ and $s^{int}_{in}(i)$ are the incoming strength on node i coming, respectively, from all the external and internal source (with $s_{in}(i) = s^{int}_{in}(i) + s^{ext}_{in}(i)$) and $s^{int}_{out}(i)$ is the outgoing (i.e., towards other pages) strength of node i. In all the mentioned cases, the strength of node i with respect to a set of pages/sources S is here defined as $s_{in(out)}(i) = \sum_{j \in S} w_{ij}$, where w_{ij} is the counts of transitions from(to) j to(from) i;
- when the walk does not stop on a node, the next node is selected among its neighbours with a probability proportional to the transition counts.

With the above defined procedure, 10^7 paths are generated from each of the 10 different external sources. The average path length is around 1.4-1.5 nodes for all the sources, with the only exception of the MainPage for which it is slightly higher (around 1.8 nodes). In Fig. 4.2, the distribution of the lengths of the simulated paths are reported, from all the available sources.

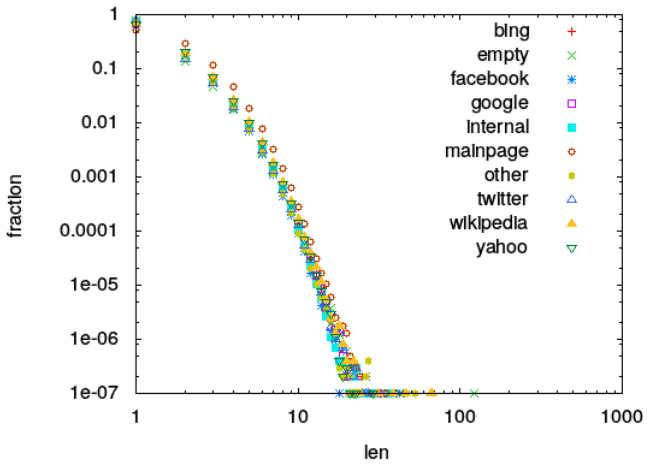


Fig. 4.2 **Distribution of the path lengths.** For the 10 different external sources, the average distributions of the lengths of the simulated walks are reported. The averages are performed over 10^7 simulations for every source.

4.1.3 Goal-oriented navigation paths

The simulated free navigation paths generated from EWC data are contrasted with real paths of users on Wikipedia, originated from a goal-oriented task. These paths are the sequence of articles followed by players of the Wikispeedia game, described in sec. 2.2.2. In particular, in the present analysis a subset of the total paths containing around 50,000 successfully paths is considered. An example of a Wikispeedia task with two different successful realizations is reported in Fig. 4.1(B).

4.2 Semantic mapping of Wikipedia pages

In order to identify whether regular patterns exist in the way the information seekers browse Wikipedia, an abstraction from the microscopic page level to a coarse-grained, semantically meaningful, representation is needed.

The problem of extracting a meaningful semantic mapping of Wikipedia pages is frequent in the literature. Many authors considered the pages content to derive a vector representation via Natural Language Processing Analysis, e.g., TF-IDF, short for term frequency-inverse document frequency, or word count analysis. This is the case of the already cited works by West et al. [111] and Ratkiewicz et al. [95]. In other studies the page content is not taken into account and one focuses instead on Wikipedia category structure. In particular, some top-level categories can be considered as main topical concepts, suitable for a semantic characterization of the pages. Here, following previous studies [70, 106, 102], the top-level subcategories of the *Main_Topic_Classification* [6] container category are treated as coordinates of a novel reduced space.

Unlike Singer et al. method [102], the semantic complexity of each page is not here reduced to just one representative topic (in their work, the one from which the shortest-path to the page is the minimal). Instead, it is assigned to each page a topic distribution, as in [70], thus mapping the article into a point of the semantic space with 13 dimensions. These 13[2] topics/main categories [3] are the very general subjects which define, as coordinates, the topical abstract space where all the pages and users' paths are studied. In the following, the coordinates are simply denotes as *topic*s.

Each article of Wikipedia is mapped into a point of this 13-dimensional space, its corresponding semantic vector being computed so that the weights are proportional to the semantic relatedness of each topic with the article's parent categories (see next paragraph). Once obtained the vectorial representation of each page, some common measures in euclidean spaces, such as norms and similarities measures, can be used to give a semantic interpretation of the position each page occupies along as their inter-relations. With these tools, the simulated paths can be read in the topical space and contrasted to some real paths, as the ones gathered with the Wikispeedia game.

Extraction of a vector representation The Wikipedia category system has a pseudo-hierarchical structure, where each page and category can have multiple parent categories. This fact, along with the lack of any central root from which the structure starts branching, turns into the possibility to quite always find a path along the categories structure to connect any category pairs. In particular this is true for

[2]On the 2015-10-22.

the main topics categories listed in the previous paragraph. Indeed, if the category tree rooted in each of the topics is evaluated, every other category can be reached via breadth-first search starting from the roots. It follows that the depth at which the category is firstly encountered in a tree is meaningful of the relevance of the rooting topic in charactering that category. Indeed, the lower the depth, the closer the category to the topic at the root, the higher the topic semantic relevance. Thus, by computing the smallest depth of each of the trees rooted in the 13 main topics, the mostly relevant topics can be assigned to each category, namely the ones from which the depth(s) is(are) minimal. The procedure followed is illustrated in Fig. 4.3 and now explained in details.

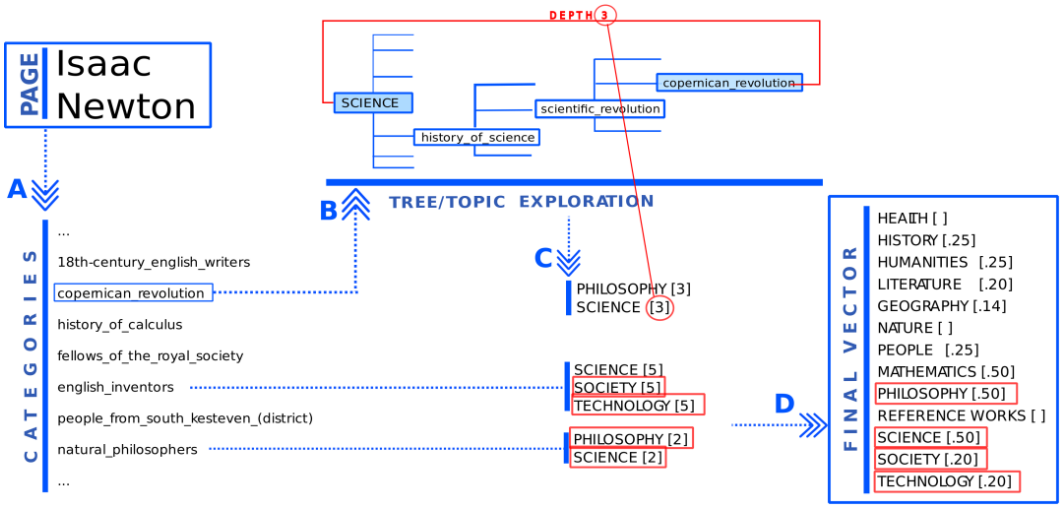


Fig. 4.3 **Example illustrating the construction of the topical vector for the Isaac Newton article.** For the Isaac Newton page one first considers the list of parents categories (panel A). For each category, one identifies the most-representative-topics (panel B), selecting the ones from which the depth of the category in the categories tree is minimal. For each page, the most-representative-topics and corresponding depths are listed (panel C). For instance the category *copernican_revolution* has the smallest depth (equal to 3) in the tree of the topic *SCIENCE*. The vector representation of the coordinates of the main topics is now obtained by weighting each topic with the inverse of the minimal depth computed above (panel D). For instance the topic *SCIENCE* appears in the topical vector with weight $1/2$.

For any page, first the parent categories which the page belongs to are listed (panel A of Fig. 4.3). To each parent category the set of the most-representative-topics is assigned. They are selected because are the ones, among the 13, from which the category depth is minimal (Fig. 4.3, panel B). In this way, one obtains for each page a set of the most-relevant-topics and their corresponding-depths (panel C). From this set, the final vector representation is easily derived by computing the weight of each

topic as the inverse of the minimal depth found for it (panel D). It has been chosen to consider the inverse so that the weights are proportional to the semantic values, and the most-representative-topics would mostly contribute in the evaluation of typical vector measures like the norm.

All the data about the category system has been extracted from the Wikipedia category links dump [5], accessed on date 2015-10-22. While generating the trees rooted in the topics, the maintenance categories are ignored, such as tracking and hidden categories. Specifically, the following categories and their direct subcategories have been filtered out: *wikipedia categorization*, *hidden categories*, *tracking categories*, *disambiguation categories*, *namespace example pages*. For some pages no vector representation could be derived, since at the time of the dump, they belonged only to some maintenance categories. They were about 5% of the total number of pages appearing in the EWC dataset and they were excluded from the successive analysis.

Robustness analysis

To validate the semantic representation obtained for each Wikipedia articles, also a different dump was considered to derived the topical space coordinates and the articles topical vectors. This dump, presented here only for robustness analysis was dated 2015-03-04. Moreover, the procedure implemented on this dataset to build the vector representation is a slightly modified version of the procedure previously described.

In the category system dump here considered the subcategories of the Main_Topic_Classification were 38, and namely: *agriculture, architecture, arts, chronology, creativity, culture, education, employment, energy, environment, geography, goods, government, health, history, humanities, humans, industry, information, knowledge, language, law, mathematics, medicine, mind, nature, objects, people, politics, science, sports, structure, systems, technology, telecommunications, universe, world*.

For each page, in extracting its vector representation based on the 38 coordinated listed above, the first three phases as explained in the main text (Fig. 2 A-C) were similarly implemented: the categories to which each page belongs to were found, and for each category its most representative topic(s) is(are) identified. It(they) was(were)

the one(s), among the 38, from which the category depth is minimal. This depth is the semantic representativeness of the topic. In the original procedure, for each topic only the smallest depth over the categories was considered when deriving the final vector. Its inverse was chosen as corresponding weight. Here, instead of considering the smallest contribution, for each topic the depths over all the categories for which that topic is the most representative are averaged. The inverse of the average is the novel weight for the topic in the final vector.

With this choice of vector representation, some of the analysis to be presented in the next sections were replicated. Since no meaningful differences emerged with respect to the main procedure presented, they are skipped from the next results description.

4.2.1 Observables used for analysis

To characterize the Wikipedia articles in the reduced semantic representation, two main measures are considered: the norm and the entropy of the corresponding topical vectors.

- The **vector norm** is the usual L2 norm, normalized to the square root of the space dimension, i.e., the number of topics. For a generic page A, whose vector w_A has components w_A^t for the different topic $t \in [1,T]$, with $T = 13$:

$$\|w_A\| = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(w_A^t)^2} \tag{4.2}$$

  With this choice, the norm is always in the $(0:1]$ range, with higher values corresponding to pages with more abstract content and lower values to more specialized pages.

- In order to measure the level of multidisciplinarity of a vector, the **entropy** $S(w_A)$ of vector w_A is computed as:

$$S(w_A) = -\frac{1}{log_2(T)}\sum_{t=1}^{T}\hat{w}_A^t \log_2(\hat{w}_A^t) \tag{4.3}$$

with $\hat{w}_A^t = w_A / \sum_{t=1}^{T} w_A^t$, so that the weights sum to 1. High values in entropy means a very general or multidisciplinary content, while the low-entropy pages are pages semantically connected with only one knowledge field.

Fig. 4.4 reports the distributions of the norms and entropies for all the pages considered as well as some example of pages lying at the extremes of the distributions. It is worth noting the particular shape of the resulting entropy distribution, reported in the subfigure on the right. Indeed, the topical vector entropies tend to distribute in isolated spikes. This is due to the input points. In fact, the norm function is defined on the the hypersquare $\mathscr{S} = \{\mathbf{w} \in \mathbb{Q}^T : w^t \leq 1 \forall t \in [1,T]\}$. However, only *few* points in $\mathscr{S}$ could be actual arguments of the function, namely the points whose coordinates are of the form $1/p$, where p could be any depth in the topic trees. Given the finite size of the Wikipedia system, the maximum depth at which any page is found in any topic tree is ~ 30. As a consequence, the input set is a finite subset of $\mathscr{S}$. For the entropy, the vectors in $\mathscr{S}$ are rescaled in order to satisfy the constraint on the coordinates sum. In this way, they are mapped into a finite subset of the hyperplane in $([0,1] \cap \mathbb{Q})^{13}$ defined by $\sum_{t=1}^{T} w^t = 1$. The spiked entropy distribution is a consequence of this particular sample of the function domain.

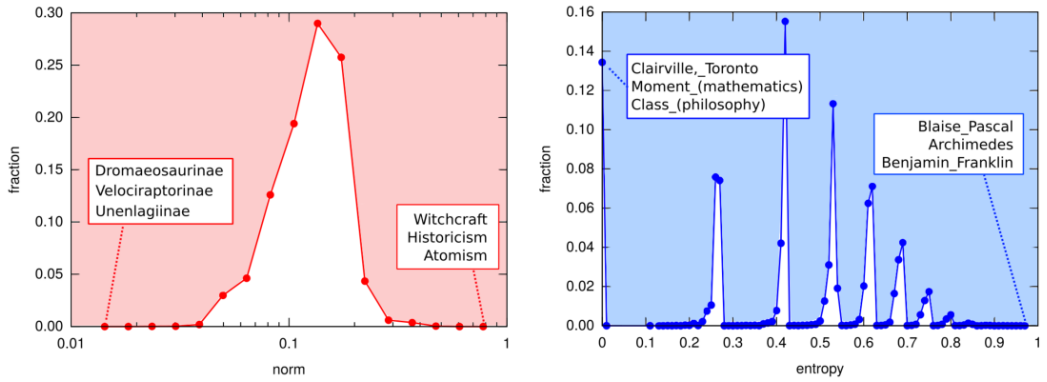


Fig. 4.4 **Distributions of page norms (left) and entropies (right).** The distributions are computed over the set of all pages for which a vector representation was derived. They correspond in the figure to the white areas under the red and blue continuous lines respectively. For both norm and entropy, in the boxes some exemplar pages are reported to illustrate the meaning of extreme values.

In addition to the observables introduced above, also distances and similarities between vectors are evaluated:

- **distance** $d(w_A, w_B)$ - It is the usual L2 distance between the vectors w_A and w_B:

$$d(w_A, w_B) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(w_A^t - w_B^t)^2}; \tag{4.4}$$

- **similarity** $sim(w_A, w_B)$ - It is the cosine similarity of the two vectors w_A and w_B:

$$sim(w_A, w_B) = \frac{1}{T}\frac{w_A \cdot w_B}{\|w_A\|\,\|w_B\|} \tag{4.5}$$

These two quantities give complementary information about how close two pages are in the semantic space. Indeed, while the distance provides an overall idea of how far two pages are in terms of both content diversity and depth, the similarity is more directly related to the extent of their semantic overlap, i.e., regardless of the difference in depths.

4.3 Results

4.3.1 Google vs Wikispeedia

This section reports the results obtained by means of the semantic measures defined in the previous section on the simulated browsing paths.

The walks have been split in classes depending on their lengths l. For each class, all the nodes encountered in any walk belonging to the class were considered and gathered according to their position k along the path counted from the end. In this way, it is replicated for sake of simplicity the same alignment proposed by West et al. in their work [111], thus assuming that the node where the navigation ends is the target node of the user surfing the encyclopedia. With this choice, in terms of notation, the first nodes encountered have index $k = l$, while the last ones have $k = 0$. Furthermore, it is denoted with w_k^l the vector of a node encountered k steps before the end on a path of length l.

At a first level of analysis, the observables defined in the previous section are evaluated by averaging over all the nodes appearing at position k of the walks, for fixed path lengths. In particular, it has been computed: (A) the average norm $\overline{\|w_k^l\|}$, (B) the average entropy $\overline{S(w_k^l)}$, (C) the average distance $\overline{d(w_k^l, w_{k-1}^l)}$ and (E)

similarity $\overline{sim(w_k^l, w_{k-1}^l)}$ between each node and the next visited along the path, (D) the average distance and (F) similarity between each node and the last one in the corresponding path, respectively $\overline{d(w_k^l, w_0^l)}$ and $\overline{sim(w_k^l, w_0^l)}$.

Figure 4.5 displays the results obtained on simulated walks where *google* was chosen as the external source to weight the starting probabilities. In this figure, as well as in the following ones, the last nodes of the paths (i.e. position $k = 0$) are aligned to the right.

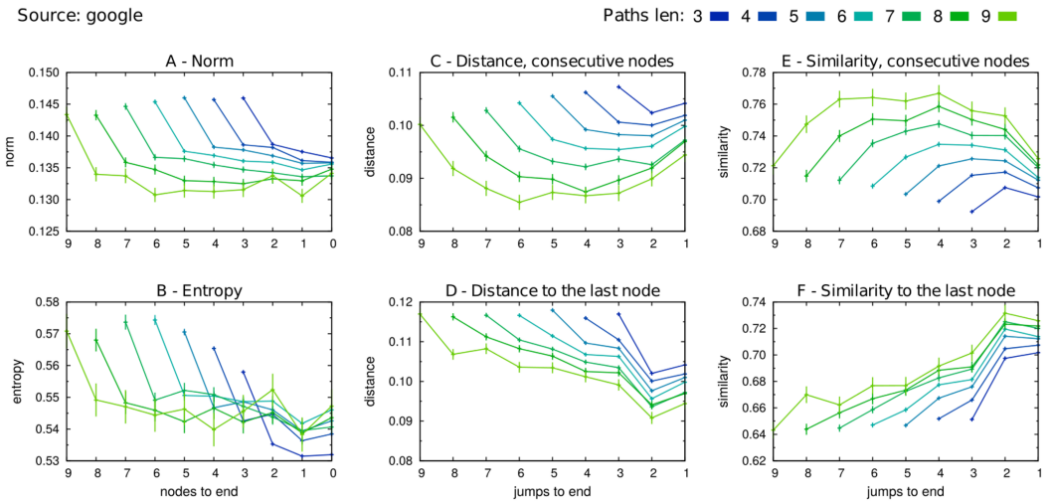


Fig. 4.5 **Paths generated from the external source *google*: averages.** The 10^7 paths simulated with *google* as source were split by lengths. For each fixed length l, the averages of the following quantities were computed over all the nodes(pairs) at k steps(jumps) from the end: (A) the average norm $\overline{\|w_k^l\|}$, (B) the entropy $\overline{S(w_k^l)}$, (C) the distance and (E) the similarity between all the pairs of nodes consecutively visited along each path, respectively $\overline{d(w_k^l, w_{k-1}^l)}$ and $\overline{sim(w_k^l, w_{k-1}^l)}$, (D) the distance and (F) the similarity between every node visited and the ending node along each path, i.e. $\overline{d(w_k^l, w_0^l)}$ and $\overline{sim(w_k^l, w_0^l)}$. The error bars display the standard errors of the means. Each color refers to a path length, from 3 (blue) to 9 (light green).

Regular patterns across different lengths of the paths can be clearly observed in the trends of the six observables. From the norm subfigure (A), it emerges that by the first step, whatever the length of the walk, the simulated walker moves from quite general and abstract pages to more specific ones, further slightly increasing the specificity of the pages while pushing the walk. Complementary, by investigating the entropy variation along the paths (B), it is suggested that the simulated Wikipedia reader typically access the encyclopedia from *google* via interdisciplinary articles, focusing on more defined field in the very first steps of her navigational path.

From the analysis of the pairs measures it appears that the reader tends to span a bigger space – both in terms of distance (C) and similarity (E) – at the beginning and ending of her walks, the steps in the middle still contributing to get her closer to the node where she stops the navigation. Quite interestingly, patterns very similar to the ones emerging in the semantic distance between consecutive nodes, subfigure (D) in Fig. 4.5 are found by Mastroianni et al. in a recent contribution [76], while investigating via GPS tracking the average spacial distance travelled by single vehicles. In their work, they found a rescaling law of the patterns, which turns into the emergence of a universal behaviour of the drivers, independent of the path length, with the longer distances between successive stops appearing – as in the present case – at the beginning and ending of the paths.

To gain a deeper understanding of the results shown in Fig. 4.5, other two sets of data are also considered. The first one is a *null model* set of 10^7 paths, still generated using *google* as external source, but after renaming the nodes, i.e. after reassigning randomly the semantic vector representations to the whole set of pages. By doing this reassignment, any semantic correlation between the pages are destroyed, while the topology of the transition counts graph is preserved. The second dataset considered is the dataset of the paths gamed in Wikispeedia [112].

Moreover, in order to investigate whether the pattern emerging are actually independent of the paths length, all the computed quantities are suitably rescaled. To this end, the data regarding each observable are normalized with the average of the considered observable over all the nodes in the paths of corresponding length. For instance, the rescaled average norm reads $\overline{\left\|w_k^l\right\|}/\langle\overline{\left\|w_k^l\right\|}\rangle_k$, with $\langle\overline{\left\|w_k^l\right\|}\rangle_k = \frac{1}{k}\sum_k \overline{\left\|w_k^l\right\|}$, and similarly for the other observables.

The global averages used to rescaled the data are here listed and shown in Fig.4.6 for the three datasets (*google*, the *null model*, and Wikispeedia): (A) the average norm $\langle\overline{\left\|w_k^l\right\|}\rangle_k$, (B) the entropy $\langle\overline{S(w_k^l)}\rangle_k$, (C) the distance and (E) similarity between consecutive nodes, respectively $\langle\overline{d(w_k^l, w_{k-1}^l)}\rangle_k$ and $\langle\overline{sim(w_k^l, w_{k-1}^l)}\rangle_k$, and finally (D) the average distance and (F) similarity to the last node of each path, respectively $\langle\overline{d(w_k^l, w_0^l)}\rangle_k$ and $\langle\overline{sim(w_k^l, w_0^l)}\rangle_k$.

As expected, no patterns emerge in the results for the *null model*. Diversely, the averages over *google* paths seems to follow a trend, e.g. the average norm or the distance between consecutive nodes decrease over paths of increasing length.

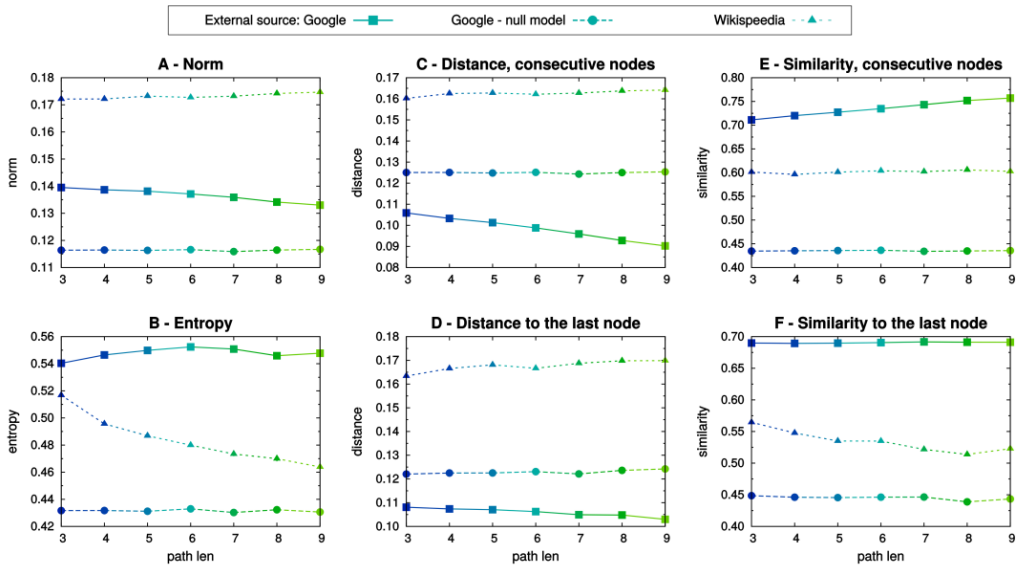


Fig. 4.6 **Averages over the aggregated paths.** The averages of measures already introduced in Fig.4.5 are here computed over all the nodes encountered along walks of fixed length l. Squares, circles and triangles refer respectively to the paths generated from the source *google*, to the same paths but semantically reshuffled (null model) and on the paths generated in the Wikispeedia game. Using the same palette of Fig.4.5, each color refers to a length. The standard errors of the means are reported, though not visible at this scale.

With regards to the scaled observables, they are reported in Fig 4.7, after having rescaled also the x-axis, in such a way that all the paths show unitary length. With this rescaling of both axes, interesting overlaps of the curves of corresponding to different path lengths are found.

First, it is worth remarking that after breaking the semantic correlations between the pages visited (in the *null model*), any regular pattern in the quantities observed disappears, as can be seen by comparing the figure in the left column with the central ones. It could also be noticed a regularity which is not destroyed in the null model, related to the presence of loops in which users jump from the last visited page to the previous visited one, before definitively ending in the last. In those cases, both in the actual paths and in the reshuffled ones, the distance and similarity between the third to last and the last nodes (which coincide) are respectively zero and 1, reflecting in the odd behaviour of the point before the last in graphs (D) and (F), central column (those cases occur in our simulations with frequency within 0.05 (for 3 jumps walk) and 0.07 (for length 9)).

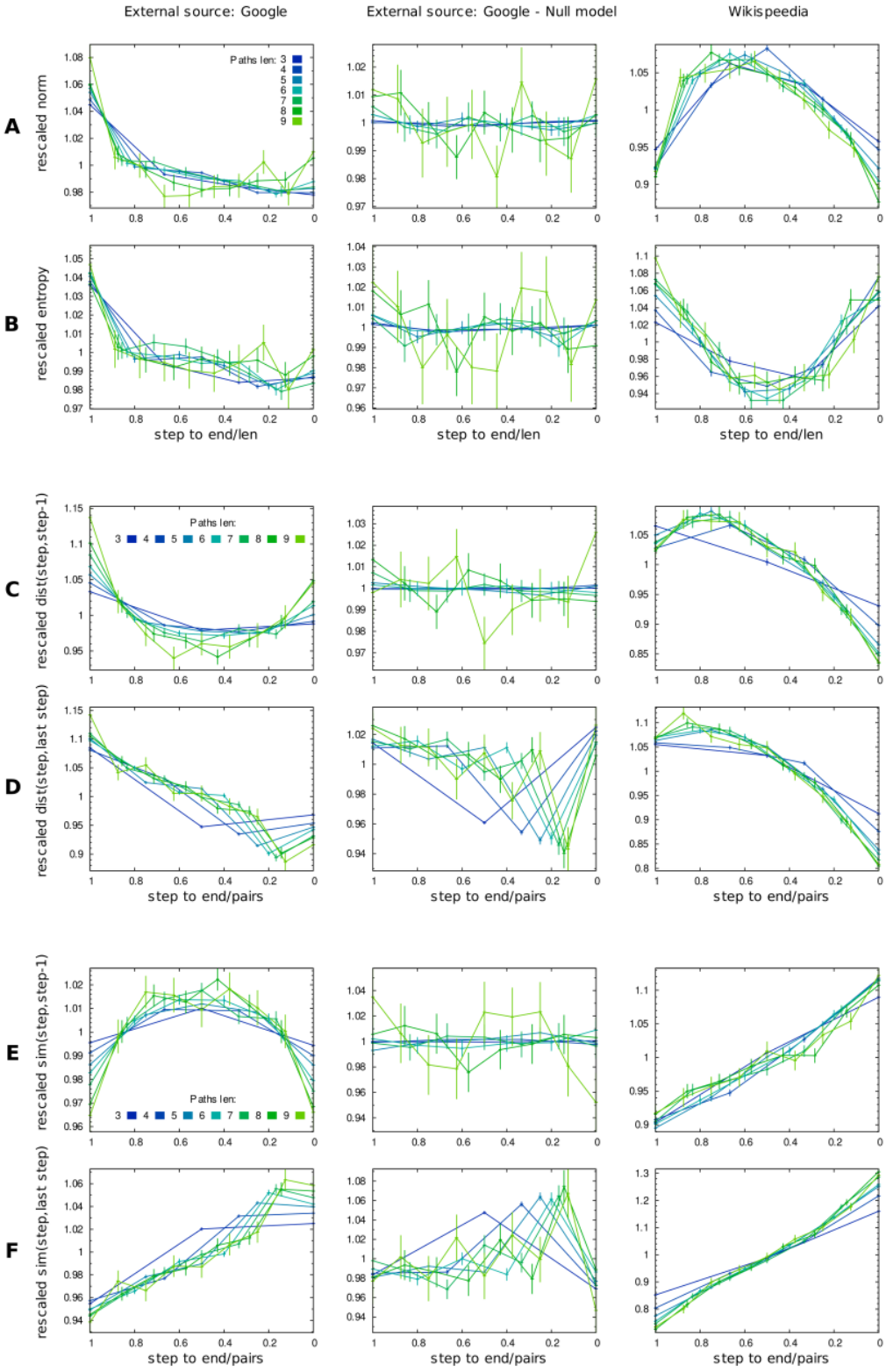


Fig. 4.7 **Rescaled averages over the simulated paths.** Continued in the next page.

Fig. 4.7 **Rescaled averages over the simulated paths.** In this panel the same data of Fig.4.5 (left column) are reported, after rescaling. The walks lengths are normalized to 1. The corresponding averages for step of the different measures (A)-(F) are rescaled with the mean value of the same measures evaluated over the whole set of nodes belonging to paths with the same length. The averages used to rescale the data are displayed in Fig.4.6. In the central and right columns similarly processed data are reported which refer respectively to a semantically uncorrelated model based on the *google* paths and to the Wikispeedia paths. Each color refers to a path length, from 3 (blue) to 9 (light green). The standard error of the means are reported.

In both the remaining cases, *google* and Wikispeedia, the curves referring to different path lengths collapse into diverse and defined patterns, thus differently characterizing the strategies followed by the users/gamers. The reader who accesses Wikipedia from *google*, enters the encyclopedia via pages more abstract and more interdisciplinary than the ones she will click on in the following steps. On Wikispeedia, where both the starting page and the target are randomly assigned by the system, the player needs to move towards most abstract articles to find her way through the hyperlinks towards the target article. This strategy is independent of the path lengths and was already observed by West et al. [111], after analysing a bunch of heterogeneous measures over the same paths. Interestingly, while the starting and final pages are quite general, the article the player goes through to connect them are more specific to narrower fields of knowledge.

Also the investigation of the pairs quantities (distance and similarity) points clearly out the different strategies of the information-seeker user of Wikipedia and of the goal-oriented player of Wikispeedia. The first steps are used by the player to make big semantic jumps, the distance to the known target starting diminishing only after some steps. From that moment on, the player gets closer and closer to the target. Still, the similarity between successive pages and between each node and the target increases monotonically. This means that every step is used to enlarge the semantic overlapping, for example in terms of common fields of knowledge supporting the corresponding vectors.

The reader of Wikipedia, instead, uses the first page to direct her navigation: the first jump is always the one connecting the most distant pages. Then the following jumps are smaller and much similar, until the last one, significantly longer than the previous ones. A similar, but reversed, behaviour emerges for the similarity.

4.3.2 Other sources of navigation

When sources different than *google* are considered, similar trends emerge, even if the entropy is in some cases less informative. In Figure 4.8 the data obtained over paths simulated from the sources *main_page* and *empty* are reported, because the mostly contrasting with what found from *google* (Fig. 4.5).

More in details, two main differences are worth being noted. First, the entropy trends are quite dissimilar. In the *empty* case it is quite flat, and no trend emerges. From the *main_page*, and unlike all the other sources, on average, whatever the path length, the entropy alternates both an increasing and a decreasing phase, as the user needed to explore diverse multidisciplinary levels in her browsing. The second worthy difference is found after contrasting norm average trends in *google* with *empty* originated paths, respectively Fig. 4.5(A) and subfigure 4.8(bA). Indeed, it is missed the large jump towards low normed articles in the first step if the user enters Wikipedia directly, from an empty referer, as she went straight to her content (and level of abstractness) of interest.

4.3.3 Measuring strategies difference

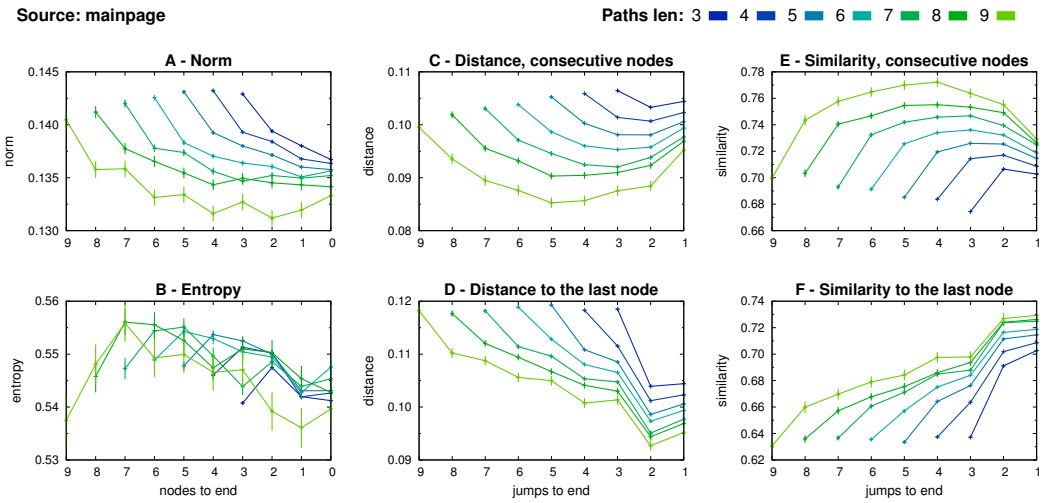
So far, the diverse trends in the observables across sources and path lengths have been only qualitatively discussed. Here, it is proposed a quantitative approach to further stress the variety of strategies found.

For the two observables norm and entropy, their unrescaled average trends along the simulated paths are considered (as the one displayed in Fig 4.5 and 4.8). More in details, for any pair of sources, say A and B, for any length l between 4 and 9, it has been computed the Spearman coefficient of the average observable along the l steps. Then, the coefficients found have been averaged over the different lengths, to synthetically express an average similarity between the two sources considered by means of a *similarity score*. For norm and entropy respectively, it reads:

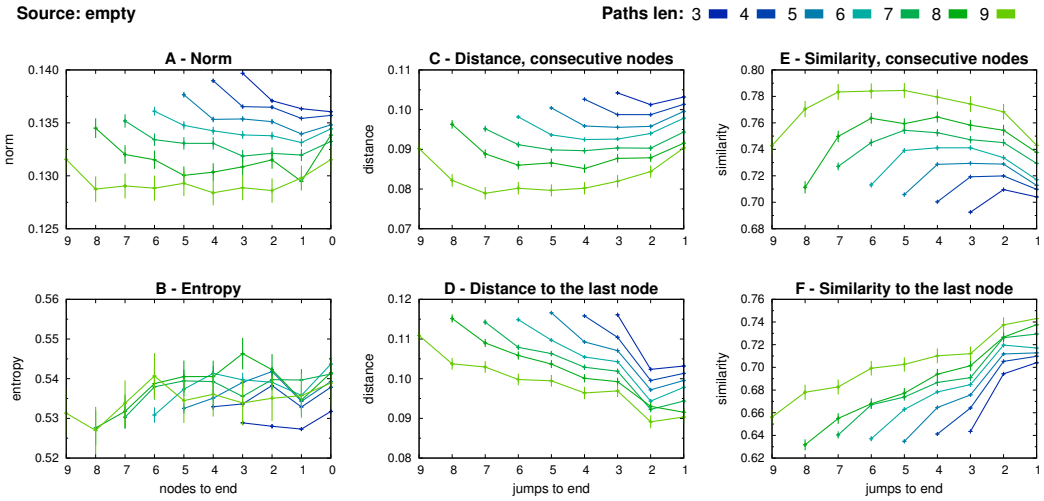
$$sim\_score(A,B)_{||} = \langle spear(\overline{||w_A^l||}, \overline{||w_B^l||})\rangle_l \tag{4.6}$$

and

$$sim\_score(A,B)_S = \langle spear(\overline{S(w_A^l)}, \overline{S(w_B^l)})\rangle_l. \tag{4.7}$$



(a) Source: *main_page*



(b) Source: *empty*

Fig. 4.8 **Paths generated from the external source *main_page* and *empty*: averages.** The 10^7 paths simulated with the two sources were split by lengths. For each fixed length l, the averages of the following quantities were computed over all the nodes(pairs) at k steps(jumps) from the end: (A) the average norm $\overline{||w_k^l||}$, (B) the entropy $\overline{S(w_k^l)}$, (C) the distance and (E) the similarity between all the pairs of nodes consecutively visited along each path, respectively $\overline{d(w_k^l, w_{k-1}^l)}$ and $\overline{sim(w_k^l, w_{k-1}^l)}$, (D) the distance and (F) the similarity between every node visited and the ending node along each path, i.e. $\overline{d(w_k^l, w_0^l)}$ and $\overline{sim(w_k^l, w_0^l)}$. The error bars display the standard errors of the means. Each color refers to a path length, from 3 (blue) to 9 (light green).

Given these definitions, any information about variability of the averages values between which the correlation is computed is lost. Nevertheless, similarities in the norm and entropy trends along any walk can be outlined, as shown in Fig.4.9. Here the similarity coefficients are reported for all the possible pairs of sources and Wikispeedia, for the norm (subplot on the left) and the entropy (subplot on the right).

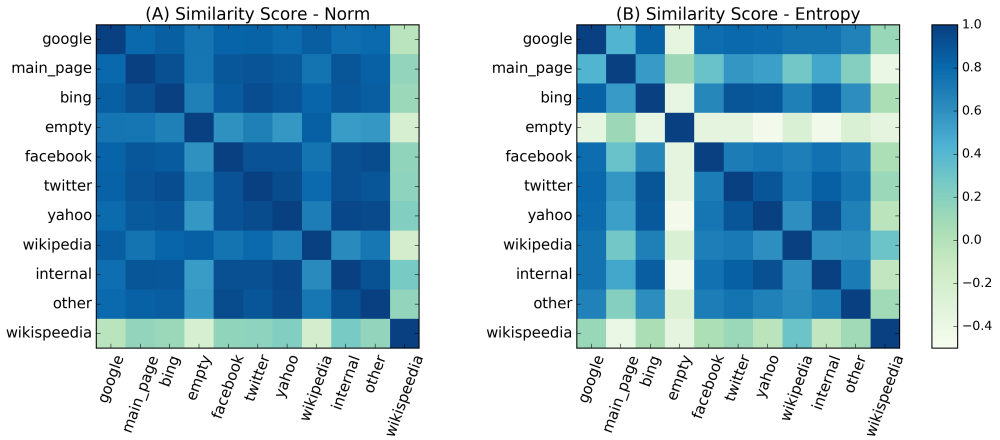


Fig. 4.9 **Similarity scores between sources.** For the two observables norm (left panel) and entropy (right panel), we report the matrix of similarities score between all the sources and Wikispeedia. The score is defined by equation 4.6 and eq. 4.7. For each pair of sources, the unrescaled averages values of the observable are considered (as in Fig. 4.5). Then, for each path length between 4 and 9, the spearman correlation coefficient is computed between the averaged values of the observable. The final score is the obtained after averaging over all the lengths.

The Wikispeedia case stands out clearly as very uncorrelated (or even negatively correlated) to all the other sources. Contrasting the EWC sources only, the entropy maps allows to confirm the qualitative observations done in the previous paragraph about the unlike behaviour of *empty* and *main_page* sources. In particular for the former, the dissimilarity with respect to all the other sources is even sharper than for the Wikispeedia case.

4.4 Discussion and perspectives

In this chapter, it has been presented an analysis of the possible strategies of the Wikipedia users. In particular, their walks over the online encyclopaedia have been simulated, based on the English Wikipedia Clickstream (EWC) dataset [115]: an

aggregated collection of clicks log of Wikipedia users activity during February 2015. In spite of the simplicity of the underlying assumption, i.e., that a memoryless Markov model could indeed be a good proxy for real user navigation [102], clear and different patterns emerge by analysing the paths on a *semantic* level.

Indeed, the simulated paths have not been analysed on the microscopic page level, rather in a more abstract space. The Wikipedia category system has been used to map each page into a point of a topical space, where different generalizations of the usual L2 metric have been defined to characterize the semantic profile of the articles in such space, e.g., norms, entropy, distances and similarities. These quantities are the observables considered to quantify the users' strategies while navigating Wikipedia. The novel procedure devised to create a semantic vector representation of each Wikipedia article is the first main contribution of the work presented.

The novel proposed semantic representation has been the tool to uncover strategies of the atypical Wikipedia reader. Indeed, while the simulations suggest that the typical distance travelled by a user is between 1 and 2 pages visited after accessing the encyclopaedia, the semantic analysis carried on has focused on longer paths over the graph.

Although the results obtained cannot be generalized as universal for the typical Wikipedia reader, because of the rareness of the longer paths analysed, still some interesting results have been obtained which uncover regularities and patterns across the simulated paths. To this end, the analysis has focused on contrasting paths originated from different sources of access into the system, with suitably devised null models and with the results based on the real navigation paths of players of the Wikispeedia game.

For instance it was observed that the longer the walk, the longer the user navigates deeper and deeper levels of specificity. Still, regardless of the length, the navigation strategy emerges as quite universal, with the very first page navigated being more abstract and of high level, and used by the reader to access her content of interest, typically more specific and concrete. As in real physical paths travelled by car drivers [76], the semantic distances spanned by the readers are not uniformly distributed along the jumps. Readers, when remaining in Wikipedia more than few steps, tend to perform the longest semantic jumps at the beginning and towards the end of their exploration. However, the semantic coherence keeps increasing throughout the paths. Moreover and in accordance with the expectations, the differences with

Wikispeedia paths reveal that Wikipedia readers do not have a well defined target in mind, their longer and rarer paths being no so goal-oriented as for the Wikispeedia players.

In conclusion, some features emerged as universal of the simulated Wikipedia users' navigation paths. When longer than 3 steps, different lengths of the paths correspond to different levels of specificity of the corresponding information-seeking tasks. Furthermore and most intriguingly, the strategies are independent of the task difficulty. Indeed, when the user keep surfing in Wikipedia, they go from an abstract starting page to access more specific content, whatever the external source they come from. Still, further work is needed to understand whether the semantic profile of the first page accessed from the external source could be an indicator of the user future behaviour, and in particular of her will to proceed the browsing.

Despite the simple assumptions made to generate realistic navigation paths, still the hints provided in this chapter represent important indications of the strategies used by learners/information seekers while exploring well structured knowledge spaces, as Wikipedia.

This kind of indications could provide important hints to improve the design of information networks and recommendation strategies. However, a systematic investigation of the true users habits while surfing a knowledge space would be key to the construction of more semantically effective learning paths. Because of the lack of availability of such data about real browsing histories, in the next chapter a similar question is addressed, but considering as navigated paths the ones suggested by textbooks authors in their pieces.

Chapter 5

Suggested learning paths: textbooks analysis

In the previous chapter, walks over Wikipedia of information seekers have been simulated based on the collection of page-to-page clickstream logs, as provided by the Wikimedia community [115]. To analyse the simulated paths, a procedure has been devised by which any article of Wikipedia is mapped into a point of a multidimensional topic space, whose coordinates are the *Main_Topic_Classification* Wikipedia categories [6]. Then, by investigating the simulated paths on this abstract semantic space and by contrasting them with real paths originated by goal-oriented tasks, some clear strategies emerged thus allowing to differentiate between users' tasks and sources of navigation.

Here, the analysis proposed aims at statistically investigating possible patterns and signatures in a very different type of *learning paths*, namely some advanced scientific textbooks. Indeed, the main assumption made is the following: every textbook is one of the possible diverse exploration paths a reader can do in the subject-related *knowledge space*. More in details, it is the path suggested by the textbook author through chapters, sections and paragraphs. While reading them, i.e., while exploring the space so-well-known by the author, the reader learns novel concepts, often revisits others and sometimes makes semantic jumps, thus creating novel connections in her proper information space, the one she is enlarging while studying the textbook.

In this perspective, any textbook study is a matter of innovations for the reader, while the author covers only a small area of the entire knowledge space. In particular, only that area the reader is ready to explore. This kind of dynamics is reminiscent of the *adjacent possible theory*, introduced by Kauffman [68] and recently reconsidered in the investigation of innovation dynamics [108, 54, 81].

Given this framework, the textbook analysis here conducted has a twofold scope. First, it aims at identifying and extracting the underlying knowledge space over which the narration walks are assumed to take place. This turns out into the problem of identifying the *knowledge units* and then the connections among them. The second goal of the proposed work is to investigate each textbook as a dynamical process of exploration of the defined spaces.

In the following sections, firstly the set of textbooks used for analysis is described, along with the preprocessing and cleaning procedures implemented. Then, two approaches are proposed to tackle the problem of the knowledge space definition. While the first takes advantages of only internal information, the second approach relies upon the Wikipedia graph as external knowledge space and on the recent software TAGME [44] to map the textbooks into a proper set of Wikipedia articles.

After defying a possible knowledge space, either a more static analysis is performed over the *knowledge* units defined and a dynamical one, thus following both the more traditional quantitative linguistic approach and the more recent one, from innovation dynamics analysis.

5.1 Textbooks: data and preliminaries

The dataset under analysis is a collection of 78 textbooks covering broad different subjects, mainly in the context of advanced Physics and Mathematics.

The textbooks were downloaded from Project Gutenberg (www.gutenberg.org) and received as raw plain text[1]. Their are listed in Fig. 5.5, clustered by their semantic interdistances, as defined in the section 5.2.2).

Every textbook was cleaned and preprocessed as described in the following paragraph.

[1]Courtesy of Stefan Thurner and Bo Liu.

5.1.1 Cleaning and preprocessing

From each textbook, only the core sections from the beginning of the first chapter until the end of the last one were preserved, thus discarding possible indexes, prefaces and appendixes. Headers and footers were filtered out, as well as figures and tables captions. The subjects indexes were also separately saved for subsequent analyses.

The cleaned texts were split into their shortest meaningful linguistic units, namely the sentences. This was accomplished automatically by means of the *PunktSentence-Tokenizer* available in the Natural Language Toolkit platform [19]. Such module is equipped with a customizable set of parameters, used to distinguish true sentence separators from functional expressions (e.g. abbreviations, unit of measures). Although such set of parameters was enriched to deal with context-typical expressions as identified by inspection, few sentences were not correctly tokenized. They could for example result in empty tokens of meaningless punctuation as well as in very long tokens, encompassing several actual phrases, or in tokens embracing both subsection titles and their next sentences. In the following analysis, this first source of noise has been taken into account.

5.2 Building the knowledge space: units

The fundamental assumption under the analysis here presented is the possibility to statistically read and study each textbook on a more abstract level with respect to the words level. To define and characterize such abstract *knowledge space*, the elementary bricks of the space must be chosen, as representative of the text content. In the following, such fundamental semantic elements are referred to as *semantic units*.

However, the units choice is not unique for each textbook. Indeed, many different sets of units could be selected as key to identify its content. Some starting attempts were performed to select the units via keywords extraction algorithms, as the ones proposed by Najafi et al. [82] or in [75]. While this keyword extraction approaches allowed to consider every word appearing in the text as a possible unit, they failed in easily identify more significant n-grams of words.

As a consequence, other two strategies have been devised and implemented to extract the *knowledge units*, leading to the definition of two different types of units,

concepts and *tags*. The first strategy is based on the author's choice of the main *subjects* covered in the textbook: the subject index. From it the units concepts are extracted and thus located in the textbook sentences. The second novel strategy has been based on the use of TAGME [44]. It is a system which is able to tag short text fragments by Wikipedia pages. In the present framework, the original sentences are the textual inputs to be tagged, while the resulting *tags* are the semantic units. Both the strategies are described in the following sections.

5.2.1 Concepts from subject index

The subject index of any book is one among the tools provided to the readers to understand the main subjects covered and locate them across the pages. Thus, it supplies the following two key information: the set of concepts (single words or n-grams) semantically representative of the textbook according to its own author and the location(s) where such concepts are mainly discussed. As a consequence, the greatest pro in referring to the subject index concepts as knowledge units is their high significance.

Nevertheless, three major weaknesses are to be considered, which could become particularly severe if one would generalize the strategy to other frameworks. First of all, and more obviously, the possibility of extracting semantic concepts from the subject index depends on the actual existence of a subject index. Thus, the analysis here performed is limited to this kind of books.

The second disadvantage is related to the difficulty of automatizing the concepts extraction from the subjects index. Indeed, different subjects indexes use different formatting rules to list words and their combinations into concepts, where here a *concept* is every instance to which a (set of) page number(s) is associated. As a consequence, for every textbook a novel set of rules should be devised to account for its particular formatting style or a by hand cleaning of the indexes must be carried on. In this analysis, the recombination of the original subject indexes entries has been performed by hand, because of the scarceness of the sample under consideration.

The final issue regards the localization of the extracted concepts across the sentences. Indeed, while the author provides the pages where the concepts are discussed, in the textbook cleaning procedure the page numbers could be filtered out, if even present in the original raw text, thus losing the possibility to directly take

advantage of the author's indications. On the other side, it is rare that any concepts composed of more than one word appears in the text exactly as it appears in the subject index, because of the words' order or their particular inflection.

To overcome this problem, the following procedure has been designed. Each concept was reduced to a set of stemmed words, after removing stop words and numbers. The same procedure was applied to each sentence, thus maintaining the set of significant words, but regardless of their order of appearance in the phrases previously selected. Then, the intersection of each possible combination sentence/concept was evaluated, thus resulting in the list of sentences where each concept was contained.

5.2.2 Wikipedia tags from TAGME

To identify the semantic units relevant for each textbook, the second strategy devised relies on an external platform: TAGME. This software system was developed by Ferragina et al. [44] and aims at annotating short text fragments by articles of Wikipedia, by taking advantages of the entire semantic context of the fragment. Before entering more deeply into the details of TAGME, some remarks are needed over the key positive and negative aspects of using this system to extract semantic significant tags.

Actually, many advantages can be identified. First of all, TAGME can be used against any textual source, without any limitation (except for the length, being the algorithm optimized to tackle short fragments). A second pro regards the outputs of the engine. Indeed, the tagging procedure results into a set of uniquely identified Wikipedia pages, together with several parameters useful to evaluate the tagging goodness. These pages are well defined and they can be broadly further characterized, since they are part of a larger system, namely Wikipedia. For instance, a semantic characterization of the articles can be made, by taking advantage of the Wikipedia category system, thus reproducing the same semantic reduction and abstraction procedure described in the previous chapter.

Concerning the most critical aspect in dealing with TAGME, the tags coherence with respect to the starting sentences must be taken into consideration. Indeed, as already stated, the input sentences could be too long, encompassing more original phrases or they can be composed by only few words, because of a not correct

tokenizing, thus turning out in ambiguous or misleading textual fragments. To limit possible incorrect tagging, the software free parameters were tuned appropriately. The final choice is described and justified in the next paragraph, along with a general description of the the software algorithm.

Software description

Details about the design of the software are reported in the original paper of Ferragina et al. [44]. This article is also referred for the terminology and the notation here used. In this paragraph, hints are provided on some technical aspects and the general idea implemented in the software.

The version of TAGME used in the present analysis was built upon the Wikipedia dump of April 7, 2016. From this, the software indexes the set of *anchors*, namely the texts used in any page whichs point to another Wikipedia page. With this, given a text fragment T, TAGME looks for all the possible anchors occurring in it. For each of them the software disambiguates their best sense, i.e. the pages to which they would more likely point given the context, to then eventually prune them, in this way discarding any not meaningful anchor.

Concerning the disambiguation phase, two features are considered to select the best candidate annotation p_a among all the possible pages pointed by any anchor a, i.e. among all its senses, $Pg(a)$. The two features are the *goodness* of the annotation and its *commonness*. The *goodness* is quantified by the score $rel_a(p_a)$. It is computed by considering all the other anchors b_i appearing in the text together with all their possible senses $Pg(b_i)$ and evaluating the relatedness of each pair of senses p_a, p_b, with $p_b \in Pg(b_i)$. After the goodness score is computed for all the senses of anchor a, only the top-ε senses are saved. Among them, the one with the highest commonness is selected as candidate, being the commonness the probability that a particular occurrence of a points to that sense p_a.

The ε parameter can be chosen by the user. A higher value should be set in case of ambiguous and short texts, while it could be decreased for long texts, where the context should be more taken into account. In the present analysis, ε is set to 0.3, i.e. to the default value in the online TAGME platform.

After the senses are disambiguated for all the anchors, the resulting candidate senses are pruned. This phase takes into account the link probability of each anchor

and the coherence of its candidate sense with respect to the candidate senses of all the other anchors. This two features are combined into a score $\rho(a \mapsto p'_a)$, where $a \mapsto p'_a$ indicates the candidate annotation of anchor a with the disambiguated sense p'_a. Only the candidates with a score greater then a threshold ρ_{NA} are kept.

This second free parameter is suggested by the authors to lie in the interval $[0.1, 0.3]$. In all the present work, the threshold is set equal to 0.3, thus asking the system for the most severe pruning of the candidate senses, in the interval of feasibility.

Finally, by inspection it was observed poor coherent annotations when the anchor contained numbers. For this reason, they were discarded from the analysis.

Semantic characterization of tags

One of the most valuable advantages in tagging the textbooks sentences by Wikipedia pages is undoubtedly the possibility of exploiting the Wikipedia framework. Indeed, the entire Wikipedia graph represents a general and well-defined *knowledge space*, where to study the textbooks. Here, connections between pages are already signs of semantic relations between them, and thus meaningful in the present analysis.

In addition, the Wikipedia category system allows to move further in abstracting the semantic characterization of the pages. This was done in Chapter 4, by devising a semantic representation of the Wikipedia articles. Here, the same procedure described in section 4.2 was replicated. From the category system corresponding to the Wikipedia dump over which TAGME is built, each tag was mapped into a topic vector. The topics, i.e. the coordinates of this topic spaces, are listed in Figure 5.3.

Given this enriched semantic representation of the tags, many different information can be obtained about the textbooks, thought as streams of tags, or even just as a collections of point in the topic space. Both the interpretations are discussed in the next sections.

5.3 Mapping the textbooks in the knowledge spaces

In the previous section, the strategies to define and extract the semantic units were described. Both the procedures result in the set of units (concepts or *tags*) occurring

in any phrases, if any. In the following, the sentences which do not contain any units are discarded.

For each textbook, the average number of units co-occurring in its (not units-free) sentences is computed. They are reported, for both units type, in Figure 5.1, along with the corresponding standard deviations.

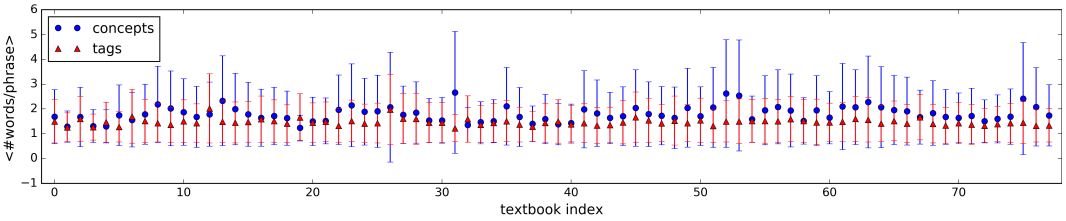


Fig. 5.1 **Average number of semantic units per not empty phrases.** For both units types, the average number of units per phrases is computed for each textbook and diplayed in the figure in function of its index. The averages are evaluated after discarding any sentence without units. The bars represent the standard deviations on the averages.

While the average number of both concepts and tags per sentence is almost in every textbook between 1 and 2, it is worth noting the presence of large deviations. Indeed, the distribution of the number of sentences as a function of the number of units occurring in them decades typically exponentially. However, in some books and in particular when the concepts are considered, the decay rate is slow enough that many more than 10 different units per phrases are still observed. Such large deviations could be mainly explained by the very different lengths of the concepts, and also by the fact that many concepts can share one of more words, thus being their co-occurrences in a sentence more likely. In addition, as already stated, the tokenizing procedure implemented cannot guarantee a precise sentence recognition.

For all these reasons, the units appearing in each textbooks were aligned, thus representing each book as a plain stream of units. More in details, from each textbook many different in line realizations were considered, to account for the possible co-occurrences of units in a sentence. In fact, the relative order of sets of units appearing in different phrases was kept unchanged, while the bunch of units co-occurring within any phrase were randomly reshuffled in each stream realization. This procedure is depicted in Figure 5.2. In the following, 100 realizations of units streams are considered for each textbook and for each unit type.

Moreover, further 100 streams are generated for each textbook and units type, where the units are completely randomly reshuffled, without any constraint. There

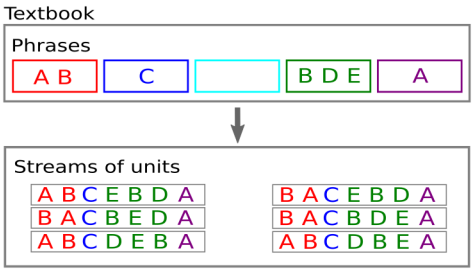


Fig. 5.2 **Textbook representation as streams of units.** Illustration of the procedure used to align the semantic units. Each capital letter represents a distinct semantic unit appearing in a sentence (coloured boxes). In order to obtain different streams of units realizations from any textbook, while the inter-phrases order of appearance is maintained, the order of any set of units co-occurring in the same phrase is reshuffled.

random streams (in the following referred to with the label *rnd*) will be considered as reference null model in the subsequent analyses.

5.3.1 Topics representation of texts

Regardless of the order with which the semantic units occur in the phrases across the text, many information can be obtained by only looking at their collection. In particular, this is the case for the *tags* units. In fact, they could be synthetically represented as points in the topic space, as described in section 5.2.2. Moreover, the number of occurrences of each tag in a text is a signature of its relevance, thus a perfect proxy for the mass of the tag in the topic space. As a consequence, each textbook can be represented as the centre of mass of the system of its tags.

More formally, given the collection of tags $\{u_i\}$ found in a textbook B and their number of occurrences n_{u_i}, the system of points $\{(\vec{w_{u_i}}, n_i u_i)\}$ is considered in the space of topics, where $\vec{w_{u_i}}$ is the topic vector of tag u_i:

$$\vec{w_{u_i}} = (w^0_{u_i}, w^1_{u_i}, \ldots, w^t_{u_i} \ldots, w^T_{u_i}), \tag{5.1}$$

being $w^t_{u_i}$ the topic-t coordinate and $t \in [1,T]$, with $T = 14$. With this notation, the textbook vector $\vec{w_B}$ could be simply derived as centre of mass of the system $(\vec{w_{u_i}}, n_{u_i})$. Its coordinates are the convex combination:

$$w^t_B = \frac{1}{\sum_i n_{u_i}} \sum_i w^t_{u_i} n_{u_i}. \tag{5.2}$$

The final vector $\vec{w}_B$ places the book B in the topic space, thus giving an immediate reduced representation of its content by the main categories used as topics. In Figure 5.3, the topic vectors obtained for two textbooks among the ones under analysis are shown.

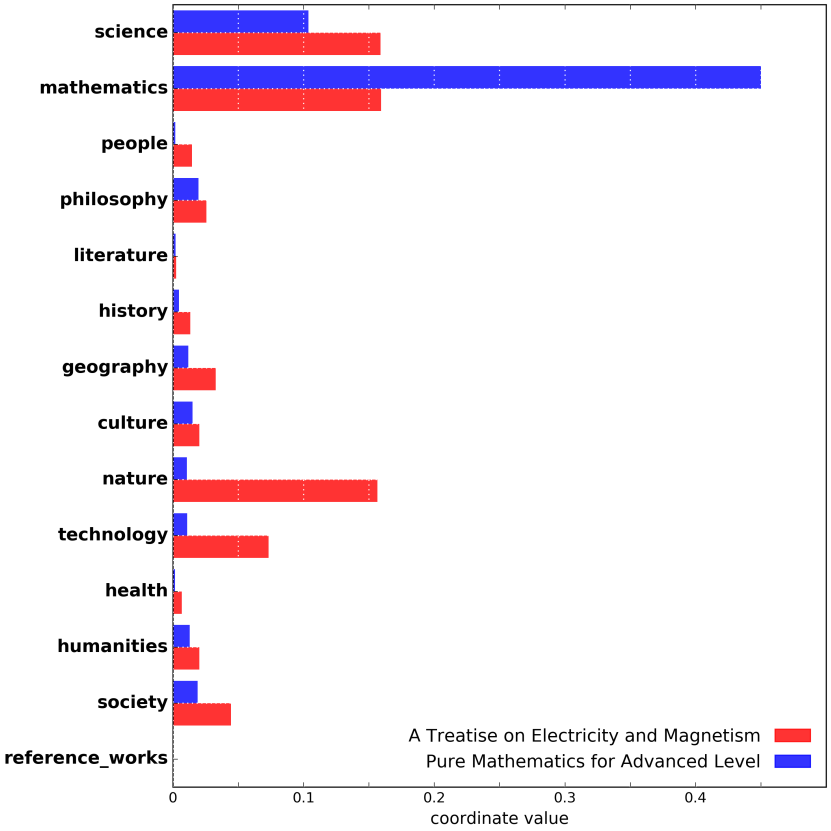


Fig. 5.3 **Textbooks topic vectors.** For each topic of the topic space, the corresponding coordinates are reported of the final topic vectors obtained for the textbooks *Pure Mathematics for Advanced Level*(blue bars) and *A treatise on Electricity and Magnetism* (red bars). The topic vectors were computed as center of mass of the tags appearing in them.

The different contributions of the topics in characterizing the textbooks are in agreement with the expectations, given the books considered, namely a textbook on Mathematics for advance level and a treatise regarding electricity and magnetism. Not only are the topics resulting as the most representative in line with the intuition, but also their relative proportion in representing the textbook content.

Starting from these observations, the analysis performed in Chapter 4 is here replicated, in order to characterized the textbook topic vectors in terms of their norms and entropies, thus quantifying respectively their abstraction and interdisciplinarity.

The two measures, already introduced in the previous chapter, are here reported for convenience. The norm is the L^2 norm:

$$\|\vec{w}\| = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(w^t)^2} \tag{5.3}$$

while the entropy is defined as:

$$S(\vec{w}) = -\frac{1}{log_2(T)}\sum_{t=1}^{T}\hat{w}^t \log_2(\hat{w}^t_A) \tag{5.4}$$

with $\hat{w}^t = w/\sum_{t=1}^{T} w^t$, so that the weights sum to 1.

With these observables, all the textbooks under analysis could be represented in a norm/entropy plane, as reported in Figure 5.4. Here some textbooks titles are also displayed, corresponding to points with extreme values of entropy or norms, among the set.

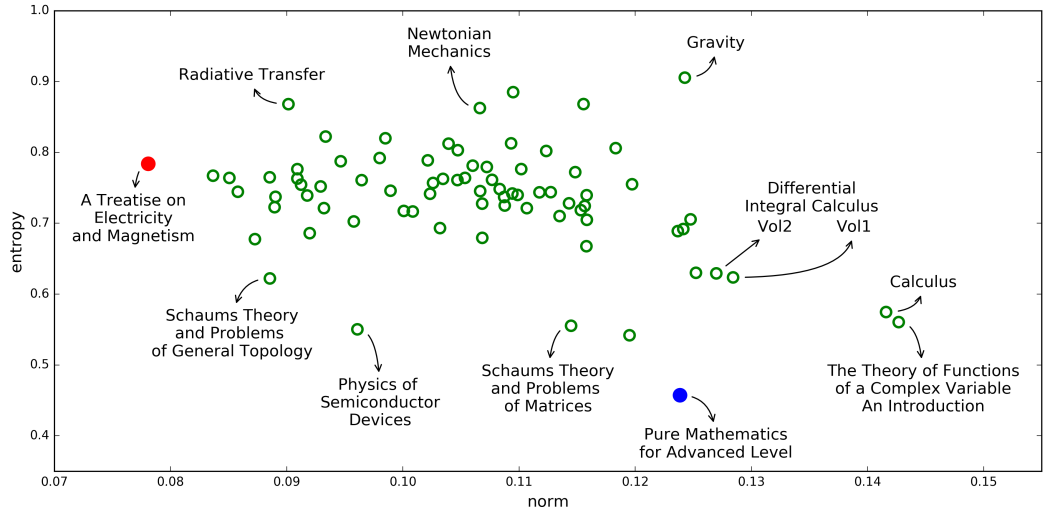


Fig. 5.4 **Norm/entropy map of the textbooks topic vectors.** For all the 78 textbooks under analysis, the norm and entropy of their topic vectors were computed. Some textbooks titles are reported as a guide to discriminate different areas of the norm/entropy plane. In blue and red the point are coloured corresponding to the textbooks whose topic vectors are displayed in Fig. 5.3.

As can be seen, entropy and norm allow to discriminate focused and abstract textbooks – as many of pure calculus and mathematics – with respect to more

concrete and topic-broad texts, such as the one concerning electricity and magnetism, highlighted in red.

In addition to the individual characterization obtained by computing norm and entropy of each textbook vector, also the semantic distances between all pairs of textbooks can be computed. The distance matrix can be the input for any clustering algorithms. In Figure 5.5 it is reported the dendrogram resulting from a hierarchical clustering of the textbooks, performed using the average linkage scheme, by means of the SciPy hierarchical clustering package [65].

The textbook hierarchical clustering based on the semantic distance between topic vectors succeeds in distinguishing two main clusters of books mainly focused in Mathematics or in Physics related issues (red and green clusters respectively. It is worth noting that books covering similar matters result to be correctly close. This is the case of the *Differential and Integral Calculus* volumes, of the college Chemistry texts and of the Electricity and Magnetisms volumes just to mention few. Future work could test the validity of the presented approach, by contrasting the semantic clustering of the textbooks with the one obtained by using more common natural language approaches, based on the *bags of words* representation of documents, or even with external, independent classification data, like the ones obtained from librarian classification systems.

5.4 Statistical signatures of the dynamics

Before completing the supposed knowledge spaces with proper connections between the *units*, here the textbooks are studied as bare streams of concepts and tags. To this end, some observables are investigated quite common in the quantitative linguistics analysis of written texts. Indeed, for decades a great effort has been made to uncover statistical regularities in human language, and to relate and explain them through generative model [100, 74, 51, 11]. In contrast with the classical approach, while more in line with very recent work [33], here the single words are not taken into consideration, but rather the previously defined semantic units. Thus, the statistical properties of the texts are investigated on a reduced, and more abstract level than the individual tokens.

Fig. 5.5 **Dendrogram of textbooks.** Given the distance matrix between any pair of textbook topic vectors, a hierarchical clustering is performed, by using the average linkage scheme. The resulting tree is displayed.

The first observable considered is the frequency of occurrences of the units into the stream, which was also historically the first to be quantitatively investigated by George Zipf in the late 30s. His empirical observation, known as Zipf's law [117], states that the words frequencies scale with their ranks r as a power-law, $f(r) \propto r^{-\alpha}$. Actually, the Zipf's law has been observed in many different complex systems, beyond the linguistic ones [74], while several refinements of the laws have been proposed to account for further empirical observations, like a double scaling in the frequency-rank plot for very large databases [47, 51].

For the five longest textbooks among the 78 under analysis, the frequency-rank plots of concepts and tags representations are reported in Fig. 5.6 Here, an exact

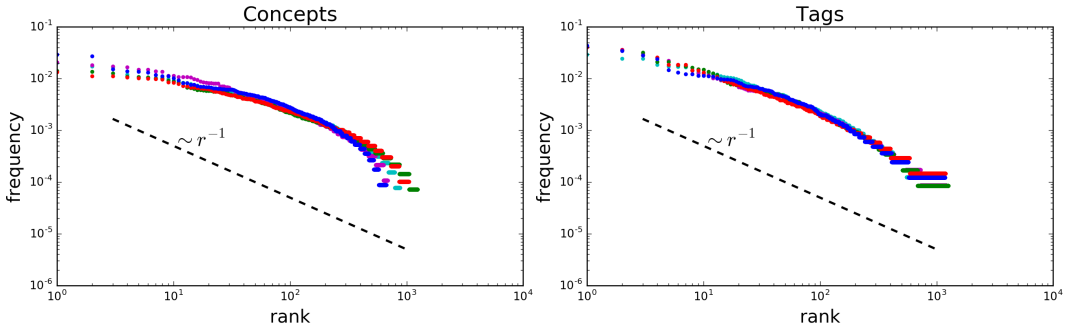


Fig. 5.6 **Zipf's law in individual textbooks.** The frequency of occurrences of the units in some books are reported as a functions of their ranks, in log-log scale. The five textbooks analysed, displayed in the plot with different colors, are the top five longest among all the 78 books of the dataset considered. The frequency/rank analysis is reported for both concepts (left figure) and tags (on the right) units. An eye-guide linear ($f(r) \propto r^{-1}$) curve is reported in both figures with dashed line.

power-law scaling is not recovered, but this result could have been expected because of the small size of the texts. In contrast, if all the textbooks in their tags stream representations are gathered together, and the frequency-rank plot is computed for all the tags appearing in the collection, a double scale power-law is recovered. It is reported in Fig. 5.7(A), along with two pure power-law eye-guide curves.

Two main observations can be made, namely the double scaling and the exponents values. With respect to the latter issue, a scaling exponent closer to 1 is usually found in texts, although large deviations have also been observed [116, 46].

Regarding the double scaling, it has been already often reveled when considering large size texts and corpora [93, 47, 51]. In particular, to explain the emergence of the two different scaling intervals, Ferrer-i-Cancho and Solé in [47] suggested that

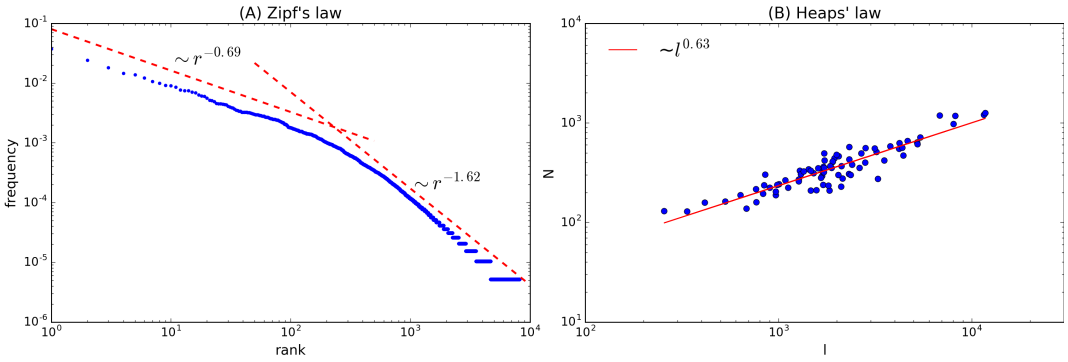


Fig. 5.7 **Zipf and Heaps' laws, tags units, entire collection.** In (A) the whole set of documents is considered as a unique text. From it, the frequency/rank plot of all units appearing is computed and reported, in log-log scale. To make more evident the double-scale of the resulting curve, two power-law of the type $f(r) \propto r^{-\alpha}$ are reported in red dashed draw, with $\alpha = 0.69$ and $\alpha = 1.62$, as resulting from a two-sloped least/squares fitting of the data. In (B), for each text of the collection, its number of distinct tags as a function of the total text length is reported. The scatters are fitted with a power-law function $N(l) = a \cdot l^{\beta}$ with exponent $\beta = 0.63 \pm 0.03$, as reported in red in the figure.

the two regimes could result from the coexistance of tho different groups of words, namely a core group of words frequent an recurrent in all textbooks (the *kernel lexicon*), and a second group of more technical and topical words (*unlimited lexicon*). While they explain the double regimes in terms of cognitive constraints of the human brain in memorizing vocabulary, recently Williams et al. [113] have proposed that the observed dual scaling is just a result of the combinations of heterogeneous texts, i.e. what they call *text mixing*.

The investigation of the frequency of occurrences of the *semantic units* provides a global insight into the textbooks. To start investigating the dynamics of exploration of the knowledge space as long as the textbook is read, a second observable is considered, namely the introduction rate of novel units in the streams. First observed by Heaps [58], the number of distinct words typically grows sublinearly with the text length, i.e., $N(l) \sim l^{\gamma}$ with $\gamma < 1$. This empirical laws has been observed together with the Zipf's one in several systems (please refer to [74] as review) and thus their coexistence has been a focus for many generative model proposed so far.

In the textbooks analysed, the sublinear growth of the number of diverse *units* with the textbooks length is recovered for both concepts and tags unit type. For the usual five longest textbooks, results are displayed in Fig. 5.8. For all the textbooks, the exponent has been fitted using a two-parameters fitting curve $N(l) = a \cdot l^{\gamma}$, via

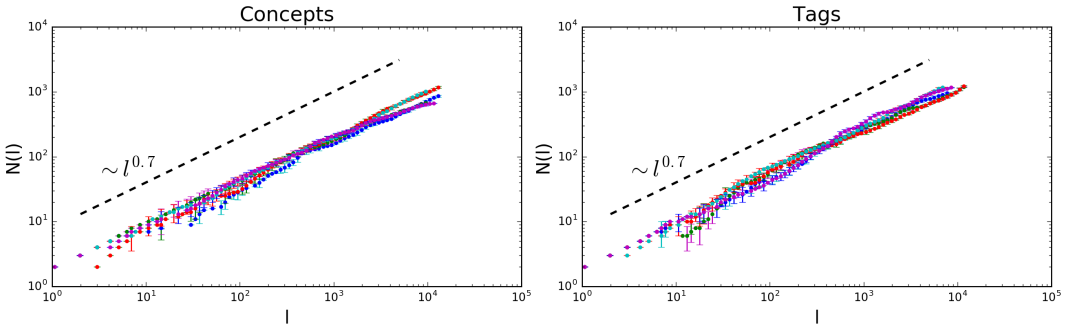


Fig. 5.8 **Heaps' law in individual textbooks.** For the same textbooks of Fig. 5.6 it is here reported the number of distinct units $N(l)$ presented as a function of the text length l, for both concepts (on the left) and tags (on the right). For each textbook, the data displayed are the averages over the 100 textbooks streams realizations. Standard deviations are also displayed. As a reference line, a sublinear curve $N(l) \propto l^{0.7}$ is reported with dashed line. Both the figures are in log-log scale.

least square fitting procedure. The resulting exponents are reported with deviations in Fig. 5.9, for both concepts and tags units (blue and red points respectively).

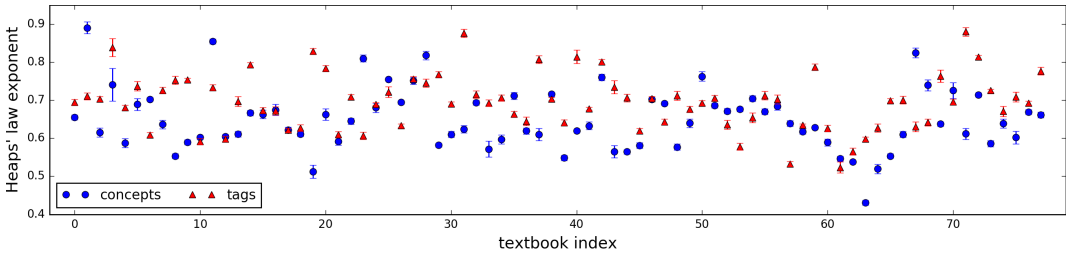


Fig. 5.9 **Heaps' law sublinear fit exponents.** For each textbook and unit type (blue and red points for concepts and *tag* respectively), the number of distinct units as a function of text length $N(l)$ (averaged over the text streams) has been fitted with a power law $N(l) = a \cdot l^{\gamma}$. The parameters were estimated by means of a least square fitting procedure. The resulting estimated exponents γ are here reported, together with the corresponding deviations.

In all cases the sublinearity is recovered, while no correlation appears between the exponents obtained in the two diverse representations. However, it is worth noting the wide range over which the fitted exponents lie, thus suggesting significant dissimilarities in the rate at which novel content is introduced among the different textbooks.

In order to consider the entire collection of texts, a diverse formulation of the Heaps' law has to be considered. In subfigure 5.7(B), for each document the size of their vocabularies are reported as a function of the total lengths, as in [93, 51, 52]. The sublinear Heapsian trend is confirmed, the vocabulary sizes growing which the

texts length according to a power-law $N(l) \propto l^{\gamma}$. The exponent γ has been estimated through a least-square fitting of the data, thus resulting in $\gamma = 0.63 \pm 0.03$.

It is worth noting that the estimated Heaps' exponent is compatible with what observed in the frequency rank plot. Indeed it is reported that, in systems fulfilling both Zipf and Heaps' laws, asymptotically $\gamma = \alpha^{-1}$ [100, 74].

The investigation of Zipf and Heaps' laws has allowed to gain insights into some statistical signature of the dynamics of exploration suggested in the textbooks. Still, other quantities can be measured to focus more on the single *unit* occurrences stream. Indeed, for each semantic *unit*, its events in a text are far from being randomly distributed. It could be expected to recover for the single *units* burst occurrences, as quite commonly observed in quantitative linguistics and social dynamics systems [67].

To look for evidence of such patterns, two quantities introduced by Tria et al. [108] are here considered, namely the entropy S of each *unit* stream of occurrences and the distribution of the triggering intervals $f(l)$. In their work, Tria et al. used the mentioned measures to quantify the presence of correlations between events associated to a same properly defined semantic group, in different systems of human activities. Among them, also some textbooks were considered, where each word was treated as constituting its own group. With this, they recovered that also in single texts, there are evidences of *triggering effects* among each word occurrences. Also in the present analysis each *unit* is treated as a single semantic group[2]. The observables definitions are here reported.

Given the stream of occurrences for each unit u across the entire textbooks, the corresponding entropy is a function of the number of occurrences of u, say k: $S_u(k)$. It is defined as

$$S_u(k) = -\sum_{i=1}^{k} \frac{f_i}{k} \log \frac{f_i}{k} \tag{5.5}$$

where f_i is the number of occurrences of *unit* u if the i-th block of the text, obtained by dividing the entire stream in k section, starting from the first occurrence of the unit. With this definition, if the k events are equally distributed over the k blocks, the total entropy is $S_u(k) = \log k$, i.e., its maximum value, while if all the occurrences

[2]However, further investigation could be done, by taking advantage of the previous defined topic vector representation of tags to create more meaningful semantic groups of words and look for correlations in the text between them.

were inside one only chunk, the resulting entropy would be zero. Finally, for each textbook, the average value is considered of the word entropies, aggregated for number of occurrences k.

Also the triggering intervals are defined from the stream of events of each *unit*. In particular, the intervals between successive occurrences are computed and thus their distribution.

To evaluate the resulting entropies and triggering intervals distributions, the same observables have been computed also on the random realizations of the original sequences.

In Fig. 5.10, for the five longest textbooks the computed entropies are displayed. As expected, in all the cases, the original sequences show lower level of entropies

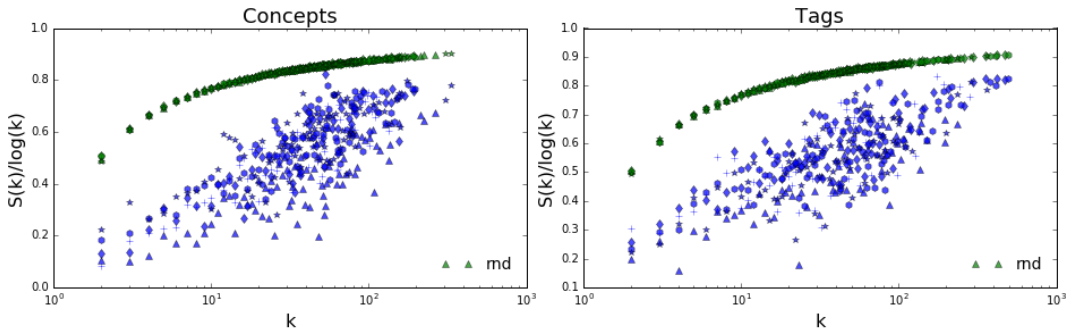


Fig. 5.10 **Normalized entropies in real and reshuffled streams.** For each textbook and unit type, the normalized entropy of the sequence of occurrences of any unit appearing was computed as a function of its number of occurrences k. The entropies displayed are the averages over all the units with the same number of occurrences k, and over all the stream realizations of each textbook. With different markers, the five different textbooks already presented in Fig. 5.6 are referred, while blue and green points concern original and reshuffled streams respectively. In both figures, the x-axis is in logarithmic scale.

for the occurrences streams, with respect to the randomized cases, in particular for low number of occurrences. The more rare a *unit* is, the more their occurrences are clustered in the text. This holds for both types of *units*.

Regarding the triggering intervals, the aggregate distributions of all the 78 textbooks are reported in Fig. 5.11, for both types of *units*. The significant presence of short intervals with respect to the random case yield a further evidence for the presence of clusters in time of the same *unit*.

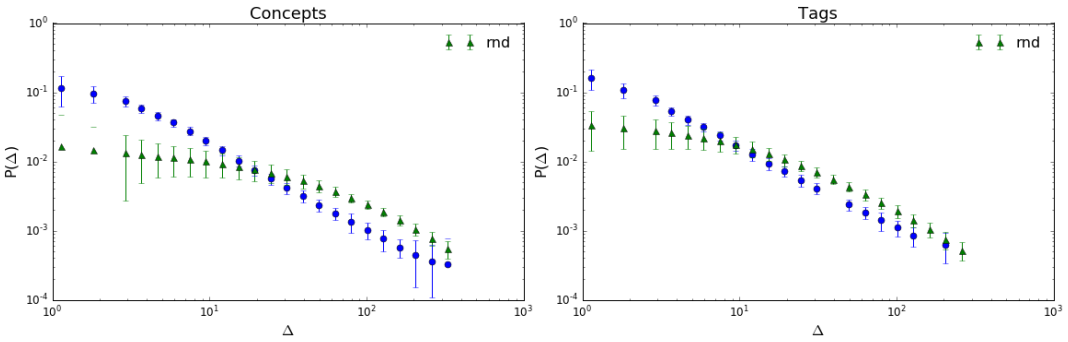


Fig. 5.11 **Average distribution of triggering intervals.** For each textbook streams and unit type, the distribution of the corresponding triggering intervals was computed, i.e. of the distances between successive occurrences of the same unit. In the figure, the average distribution computed over all 78 textbooks is displayed with blue points. The same procedure was applied to the reshuffled streams. The corresponding final average distribution is reported with green points. Standard deviations are also reported. All the axes are in log scale.

5.5 Building the knowledge space: connections

The analysis presented so far has been focused only on the stream of semantic units across the textbooks, regardless of any semantic correlation between them. However, possible correlations between the units could be relevant to better understand the statistical regularities found, as well as to quantify novel aspects of the dynamics of exploration in the knowledge spaces.

As for the choice of the elementary units constituting the spaces, several different definition of *semantic connections* between the units could also be proposed. In particular, both information grasped from the textbooks themselves and from external sources could be considered to define and quantify any semantic connection.

Here, two different approaches are presented. The first one refers to the use of the textbooks streams to quantify how much a pair of units is correlated. In particular, it is based on the entire occurrences history of all the units appearing in a single textbook. Thus, it is named here *GLocal* approach. In the second case, the source of information for inferring any correlation between units is the entire Wikipedia graph.

In the next section, the first novel approach is presented along with the drawbacks encountered in dealing with it. Then, some indications on the Wikipedia graph are reported.

5.5.1 GLocal approach

Computation of semantic correlations

The assumption underlying the approach here presented is the following: two units are semantically correlated if their temporal patterns of occurrences across a textbook are positively correlated. However, for any unit, its temporal pattern of occurrences is a plain binary stream, where only presences or absences of the units are given. Many common measures [101] of (dis)similarity between boolean vectors only compares the binary vectors bit by bit, by evaluating variables matches or mismatches. Here a different measure is needed to account for closeness, and not just overlapping, of the units occurrences.

To this end, the original sequence of occurrences of each unit across the phrases in the text was transformed into a temporal series, by mean of a running average procedure. Given the starting binary sequence for any unit $occ_u(t)$, a novel series $\mathrm{RA}_u(t')$ is built so that:

$$\mathrm{RA}_u(t') = \frac{1}{n} \sum_{t=t'-n}^{t'-1} occ_u(t). \tag{5.6}$$

The window width n was chosen differently for each textbook. It was set to be the mean inter-distance between successive repetitions of the same unit, averaged over all the units appearing in the text. An example of the transformation from the temporal binary vectors of occurrences to the smoothed temporal series for two concepts appearing in the textbook *Advanced_Level_Physics* is displayed in Fig. 5.12.

Then, for every pair of units $u1$ and $u2$ and corresponding temporal series $\mathrm{RA}_{u1}(t)$ and $\mathrm{RA}_{u2}(t)$, their Spearman correlation coefficient was computed, $\mathrm{spear}(u1,u2)$. To give significance to the resulting correlation coefficients, the following null model was designed. Given any sequence of occurrences $occ_u(t)$, 100 randomizations were performed of the original phrases/occurrences patterns so that:

- the distribution of number of units per phrases was unchanged;
- the first occurrence of each unit each was maintained.

On these partially randomized sequences, both the running averaging and the correlations procedure were performed. In particular, for every pair of units (u_1,u_2),

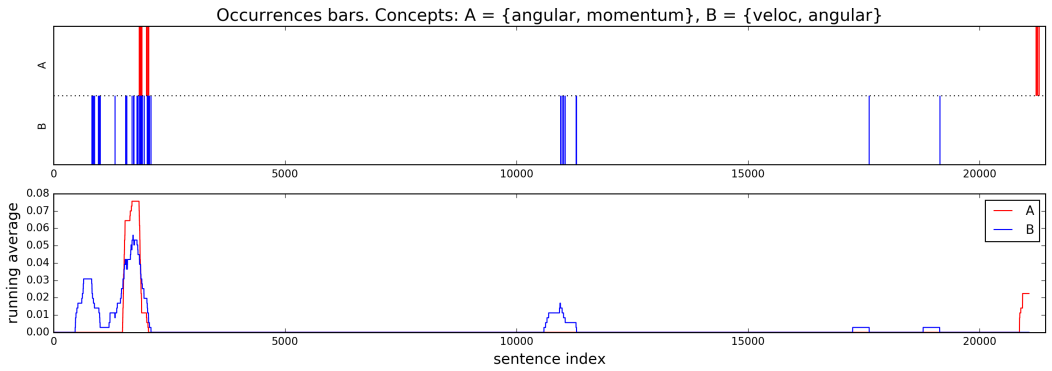


Fig. 5.12 **From binary occurrences vectors to smooth temporal series.** For the two concepts $A = \{angular, momentum\}$ and $B = \{veloc, angular\}$ appearing in the textbook *Advanced_Level_Physics*, their boolean vectors of occurrences across the textbook sentences are displayed on the top. In the figure at the bottom, the smoothed temporal series are reported, obtained after a running averaging of the occurrences vectors, as in eq. 5.6. Their resulting Spearman correlation coefficient is $\text{spear}(A,B) = 0.37$.

the correlation coefficient of their randomized sequences was averaged over all the randomizations, resulting in $\overline{\text{spear}_{rnd}(u1,u2)}$, with associated deviation $\sigma_{u1,u2}$. At the end, only the pairs of units for which

$$\text{spear}(u1,u2) \geq \max(0, \overline{\text{spear}_{rnd}(u1,u2)} + 3\sigma_{u1,u2}) \tag{5.7}$$

were considered significant. A weighted connection was drawn between them, with weight equal to their correlation coefficient.

GLocal knowledge spaces

The procedure described in the previous section allows to enrich the knowledge space of both concepts and tags with connections between the units, for every textbook in the database under analysis. In the table 5.1, some topological properties of the resulting graphs for some textbooks are reported, for both *unit* types.

In particular, only the data regarding the first five biggest graphs are reported (corresponding to the five top longest textbooks, already displayed in Fig. 5.6 and 5.8). For these five textbooks, the greatest largest component covers more than 90% of the graph. Both concepts and tags graphs have a high clustering coefficient (with respect to a random graph with similar size and edge number). Moreover, the concepts graphs show typically an assortative profile, more significant than in the

Table 5.1 **GLocal graphs properties for some textbooks.** For the textbooks listed below, some topological properties are reported of the Glocal graphs build on them for both concepts and tags units (respectively white and violet background cells). In particular, the following quantities are reported (from left to right): the number of vertices, the number of edges, the fraction of node belonging to the greatest connected component, the mean clustering coefficient, the assortativity index[84] and the mean shortest distance between pairs of nodes in GCC. Please refer to Appendix A for definitions.

Textbook id – title	#vertices	#edges	GCC	cc	a	⟨dist⟩
0 – Advanced Level Physics	877	4713	0.97	0.44	0.21	4.8
	977	6382	0.92	0.19	-0.02	4.3
12 – College Chemistry	687	3172	0.91	0.34	0.25	4.4
	1211	11460	0.97	0.25	-0.13	3.6
26 – General College Chemistry	1239	9192	0.95	0.40	0.21	4.2
	1262	11065	0.98	0.30	-0.08	3.7
46 – Physical Chemistry 4th	1046	5344	0.90	0.34	0.25	4.8
	1191	7795	0.92	0.35	0.06	4.2
48 – Physics For The Enquiring Mind	668	2550	0.91	0.27	0.09	4.5
	1177	10194	0.93	0.25	0.08	4.0

tags graphs. Finally, in both types of graph, the small world property is recovered, as the average shortest path between nodes in the greatest connected component is around 4.0.

For the same GLocal graphs, the degree distributions are displayed in Fig. 5.13. It is worth noting that the nodes in concepts graphs have not a large heterogeneity,

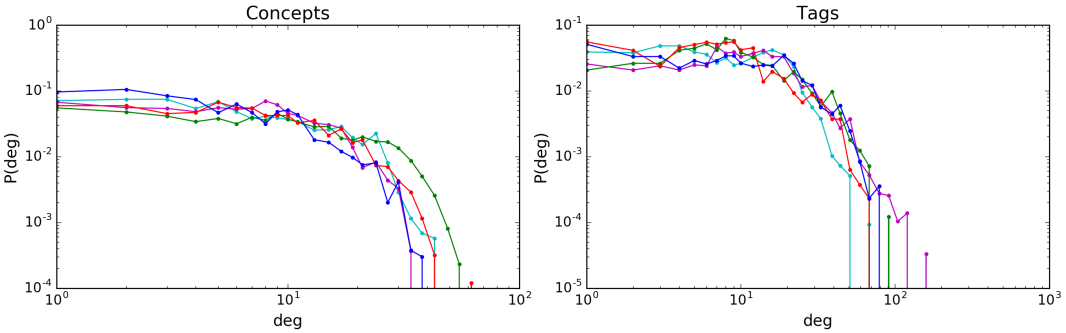


Fig. 5.13 **Degree distributions for some Glocal graphs.** For the GLocal graphs obtained from the textbooks listed in Table 5.1, the corresponding degree distributions are reported, for both concepts and tags graph units types (left and right subfigure, respectively). Axes are in logarithmic scale.

while in contrast, the degree distributions display fatter tails in the tags graphs. In both cases, the distribution is homogeneous for the smallest connectivity values (degree less than 10).

As soon as the smaller textbooks are also considered, the degree distributions become very homogeneous. The textbook size (typically around 100 units) do not allow for connectivity hubs to appear. Moreover, many disconnected components are observed, along with large fractions of completely disconnected nodes.

Because of such properties, this type of textbook network-representation seems not to be ideal for the subsequent analyses. As a consequence, only the networks over the Wikipedia global graph will be considered for further investigations, presented in the following sections. Still, the approach here presented could provide useful hints for future works on innovative network representations of books.

5.5.2 Global graph: Wikipedia

The Wikipedia graph, composed by articles connected via hyperlinks, is the straightforward structure in which the tags are naturally embedded. Furthermore, and how already discussed in the second chapter, it is the largest, collaborative, freely available realization of a *knowledge space* as a complex network.

The version here considered was build upon the dump [5] dated April 7, 2016, i.e. the one on which also TAGME [44] software version here used was based. Since redirects were not filtered out, the final directed graph contains aroud 12m articles and around 390m edges. Among all the nodes, around 15k distinct tags are counted over all 78 textbooks under analysis.

5.6 Exploration of the knowledge space

In the previous section, two different approaches were presented to enrich the knowledge spaces with connections between the semantic units. However, for the reasons already discussed, only the results obtained on the Wikipedia knowledge graph are here presented.

In particular, to characterize the dynamics of exploration on the graph suggested by each textbook, the following measures have been considered:

- **topological distance $\mathbf{G_d(u1, u2)}$** – it is the shortest path on the graph from tag $u1$ to tag $u2$. If no directed path exists from tag $u1$ to tag $u2$, their topological distance is set to $G_d(u1,u2) = -1$.;
- **topic distance $\mathbf{T_d(u1, u2)}$** – it is the euclidean distance between the topic vectors $\vec{w_{u_i}}$ corresponding to the units u_i, as introduced in sec. 5.3.1. For convenience, it is reported here its expression:

$$T_d(u1,u2) = d(\vec{w_{u1}}, \vec{w_{u2}}) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(w^t_{u1} - w^t_{u2})^2}; \tag{5.8}$$

In Fig. 5.14 for one textbook and then for the entire collection, the histogram of topological distances between consecutive pairs of tags (subfigure (A)-(C)) and between only new-introduced tags (subfigure (B)-(D)) are reported, together with the results obtained in the randomized sequences. It is worth noting that in both cases

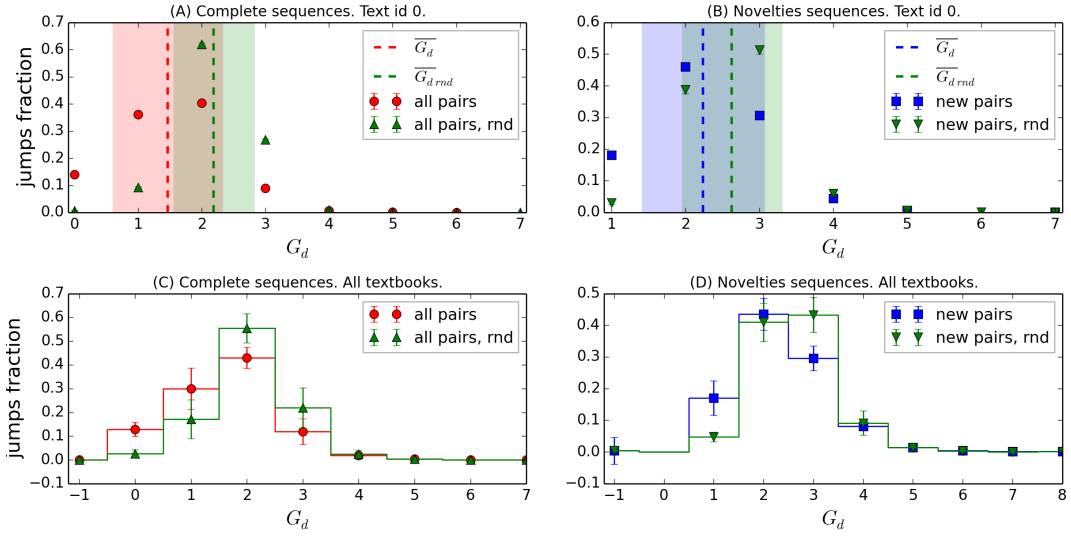


Fig. 5.14 **Topological distances between textbooks tags**. In subfigure (A), with red points the histograms of topological distances between consecutive tags in the textbook *Advanaced Level Physics* is reported. The dashed line corresponds to the average distance, while the area coloured in red defines 1 sigma of deviation from the norm. Green points, line and area refer to the randomized sequences. In subfigure (B) the same data are reported, but computed on the sequence of novel tags, i.e., discarding any repetitions. After averaging the histograms over all the textbooks, results are shown in subfigure (C) and (D), respectively for the topological jumps between pairs of tags in the complete sequences of in the novelties one. Standard deviations are reported.

(all pairs, only novel tags) shorter jumps are more recurrent in the true sequences

with respect to their randomized versions. This is more significantly true for the complete sequences, where a large fraction of jumps is made on the same node. Indeed, around %10 of times, the same tag is subsequentially repeated, thus yielding null topological distance spanned.

Similar results are obtained if the topic distance is computed instead of the topological ones. Results for the *Advanced Level Physics* textbook and then averaged over the entire collection are reported in Fig. 5.15(A)-(C) and (B)-(D), respectively for all consecutive pairs of nodes and for only novel tags pairs. For all the textbooks,

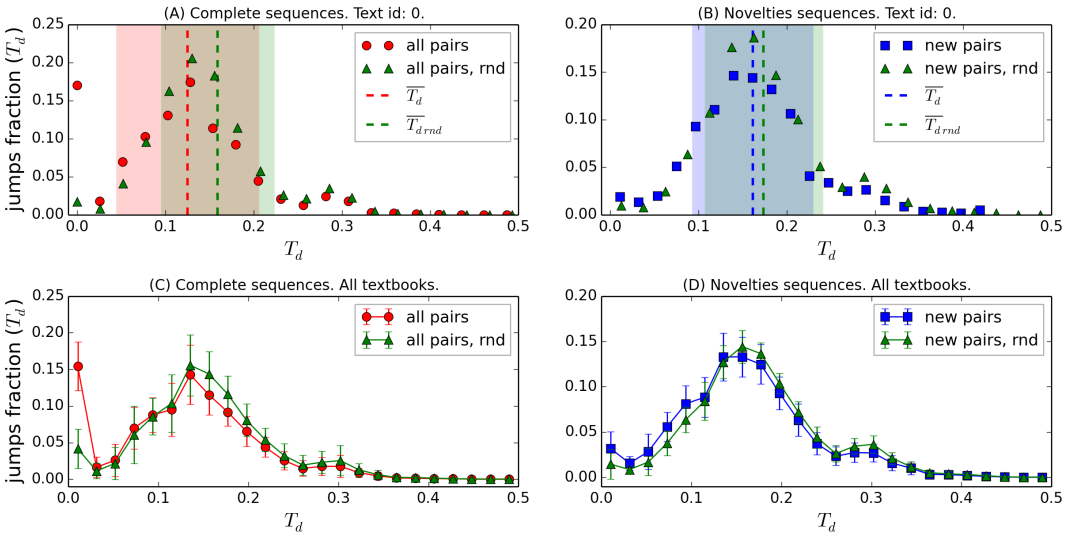


Fig. 5.15 **Topic distances between textbooks tags**. As in Fig. 5.14, in subfigure (A), with red points the histograms of topic distances between consecutive tags in the textbook *Advanaced Level Physics* is reported. The dashed line corresponds to the average distance, while the area coloured in red defines 1 sigma of deviation from the norm. Green points, line and area refer to the randomized sequences. In subfigure (B) the same data are reported, but computed on the sequence of novel tags, i.e., discarding any repetitions. For all textbooks, the binned distributions (as the one reported in subplot (A) and (B)) are averaged. The final values together with the corresponding standard deviations are displayed in subfigures (C) and (D) for respectively all consecutive pairs of tags (red points) and only between novel ones. (blue points).

their distributions are averaged and displayed in subfigures 5.15(C) and (D), together with their randomized versions. Although in all textbooks the true topic jumps are on average shorter than in the randomized case, it is to be noted that only if all the repetitions are considered a significant difference appears in the original sequences. Indeed, a peak on topic distance around zero is in the complete true

sentences always recovered, as a (partial) results of the repetitions of the same tags across the sequences.

However, to better understand the possible implications of a null topic distance on the topological characterization, correlations and overlapping of the two measures are investigated. To this end in Fig. 5.16, left subfigure, the distribution of topic distance between nearest neighbours tags on the graph is reported. On the right, the histograms of topological distances between tag pairs with null semantic distance is instead shown. Tags with the same topic representation are typically 2 or 3 steps far

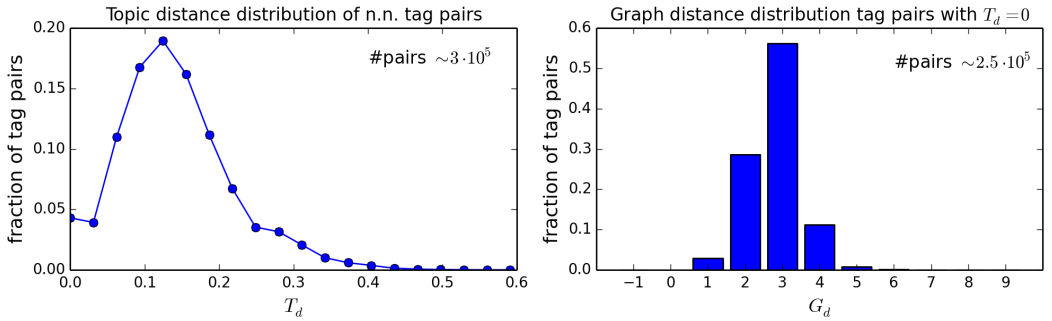


Fig. 5.16 **Topological and topic profiles of *close* tags pairs.** In the subfigure at the left, all the pairs of tags which are nearest neighbours on the Wikipedia graph are considered. Of them, the topic distances are computed. Their distributions is reported in the plot. In the subfigure at the right, first the pairs of tags with null topic distance are selected. Than, their topological distances are evaluated, and reported in the figure.

on the graphs while only a small fraction (around 5%) of nearest neighbours pairs of tags have the same semantic profile.

Moreover, it has been computed the correlation between the two distances, looking for the Spearman correlation coefficient computed over all pairs of tags appearing subsequently in at least one textbooks of the collection. It results to be $\rho = 0.26$ with p-value $= 0.$. The two distances are slightly correlated, but still the information they provide are not overlapping. Rather, it could be interesting to investigate whether the semantic distance could be useful to forecast missing links in the Wikipedia graphs, for example by looking at the semantic distances of any Wikipedia page with its topological neighbours, across its revisions history.

Still, regarding the exploration of the graph along the textbooks, the empirical distributions of jumps discussed so far can steer possible stochastic dynamics reproducing the graph exploration.

5.6.1 Expansion of the adjacent possible

The analyses proposed so far provide some insights into the dynamics of exploration of the Wikipedia space which the reader is steered to follow as she reads the books. It is understood that the suggested path is not continuous either on the Wikipedia graph or in the topic abstract space, although close tags (in both spaces) appears more frequently sequentially than in a random case.

Still, in the analysis of the spanned spaces, nothing is assumed about what area of the graph is really available. Indeed, so far the entire Wikipedia graph has been treated as always explorable by the learners, without any constraints. Actually, any learning experience is strictly bounded by what already known, while propaedeutics issues restrict the space of what can be learned as well as the mere possibility to find ways toward the unknown.

This scenario reflects, in a learning and information-retrieval context, the "*adjacent possible*" theory proposed by Kauffman [68] in a biological framework while reasoning about biosphere evolution. The adjacent possible is what is ahead of time, continuously expanding and reshaping "*at every step forward in the unknown*" [54].

Recently, the same suggestion has driven the quantitative investigation of the dynamics of innovations and creativity [108], as they appear in many social systems like social annotation of music products [81], or the network of movies and inspiration links between them [54]. In particular in the latter work, the notion of adjacent possible at each timestamp of the system evolution has been formalized by looking at its coverage, i.e. its a posteriori realization as actual history of the system.

Indeed, the main problem in dealing with the quantification of what could be realized ahead of time in any system evolution is the lack of information about its possible future states. In this sense, any evaluation of the adjacent possible can only be done *a posteriori*.

In the learning context tackled here, the stated issue about the knowledge of the future possible states of the system can be partially overcome. Here, the *adjacent learnable* can be introduced for any learner as that part of the knowledge graph – Wikipedia in the present formalization – actually accessible, for instance because their preliminary requirements have already been satisfied, or just because a bridge towards them is stated from something already known. Any novel concept learned

contributes to cover the adjacent learnable as well as to expand it, by making novel materials accessible.

Thus, if the Wikipedia graph is considered as the (more o less stable) compendium of all the knowledge available, at each timestamp of any textbook narration the entire set of nodes accessible to the next step exploration would be available. From the reader point of view, only a subset of that nodes will be learned, thus becoming part of her individual graph of knowledge.

However, how this individual knowledge graph expands in relation to the adjacent learnable can be expected to depend on the knowledge background of the reader (how much of the collective space does she already know?), author's choices and content. Does the adjacent learnable expand steady along the book or its expansion is characterized by bursts? How fast is it covered by the introduction of novel materials rather than enlarged?

In the following, some preliminary attempts are discussed aiming at formalizing the previous questions. To this end it is worth noting that, in contrast to the system of cinematographic influences between movies studied in [54], here no direct information are available about the appropriate order of exploration of the tags in the Wikipedia graph. Instead, the most simple assumption is made, namely that the edges directionality in Wikipedia constrains the space properly accessible. With this assumption, each node exploration makes its out-neighbours accessible, i.e., puts them into the adjacent learnable.

Still, two different estimates can be done. In fact, the expansion of the adjacent possible could be analysed with respect to the entire context, represented by the whole Wikipedia graph. In this case, every time a novel node is explored, all its out-neighbours became available for further exploration, thus spanning over many different semantic contents, as typically in any Wikipedia page. That would be lead to a large estimate of the actual adjacent learnable as it appears from the reader perspective.

On the other hand, the adjacent possible analysis could be narrowed to look only at the subset of tags semantically related, because they will be explored during the book narration. Rather than looking for the adjacent possible, the *adjacent future* would be in this way investigated, as a proxy of the way in which the author organize and present a predefined set of material in the narration. In the following both

the large estimate of the adjacent learnable and the adjacent future are defined and analysed.

The adjacent possible at each time t, $AP(t)$ is valued on the entire graph of Wikipedia. In particular, given its directed graph $G(N,E)$, first it is defined the set of tags already actualized at time t:

$$\mathscr{A}(t) := \{u \in G(N,E) : t_0(u) \leq t\} \tag{5.9}$$

where $t_0(u)$ denotes the first time unit u has been introduced in the unit stream. With this, the adjacent possible $AP(t)$ is defined as follows:

$$AP(t) := \{v \in G(N,E) \setminus \mathscr{A}(t) : \exists E(w,v) \in G(N,E) \text{ for some } w \in \mathscr{A}(t)\} \tag{5.10}$$

in which $E(w,v)$ is an edge from tag w to tag v. Equation 5.10 defines the adjacent possible as the set of all nodes not already actualized, which are reachable by at least an edge from the set of nodes already introduced.

As already pointed out, this definition provides a large estimate of the space actually available, without constraints on the semantic relatedness of the novel adjacent nodes.

Still, given the above definitions, several quantities can be defined to characterize how the adjacent possible grows and is covered. First of all, the rate at which it enlarges as new materials are introduced by the author is investigated. As shown in Fig. 5.17 for the usual five longest textbooks, the adjacent possible expands as a power of the number of tags introduced. There the curve are obtained by looking at the average size of $AP(t)$ averaged over all the streams realization for each textbook as a function of the average size of the actualized set $\mathscr{A}(t)$, corresponding to the number of distinct tags introduced up to time t. For smaller texts, the power-like trend is confirmed, in particular for the curve tails. The resulting exponents of least squares fitting procedures of the curves with a function of the type $AP(N) = aN^{\beta}$ are reported in Fig. 5.18. In the vast majority of cases, the adjacent possible grows sublinearly with the number of distinct tags presented in the stream. This means that, the more the reader progresses on the textbook study, the fewer the novel concepts lead to a further increase of the adjacent possible. Still, there is a large variability in the resulting exponents, thus meaning a large variability in the ways the different authors let the readers explore the space. Indeed, as an extreme exception, the

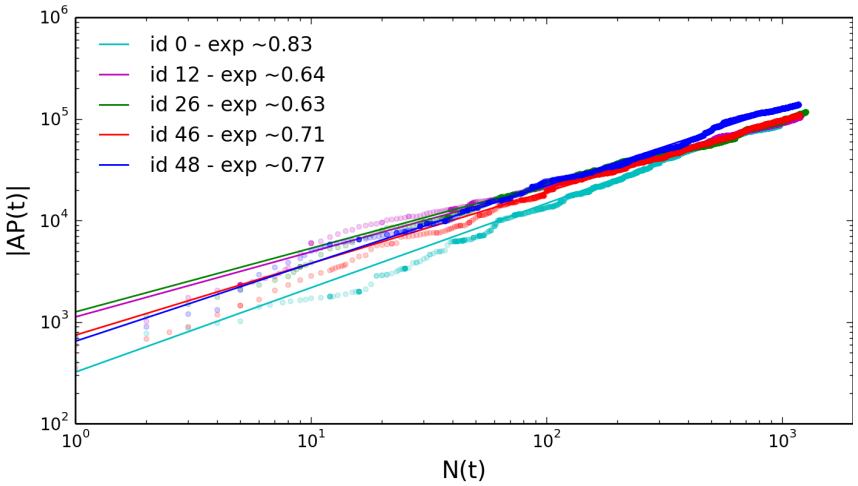


Fig. 5.17 **Adjacent possible expansion as a function of the number of tags introduced.** For the five longest textbooks, the cardinality of the adjacent possible is reported as a function of the number of distinct units introduced $N(t)$, corresponding to $|\mathscr{A}(t)|$, after averaging over the different stream realizations. For each textbook, the average curve is fitted via a power-law function $AP(N) = aN^{\beta}$, via least square method. The resulting fitted curves as well as their exponents are also displayed.

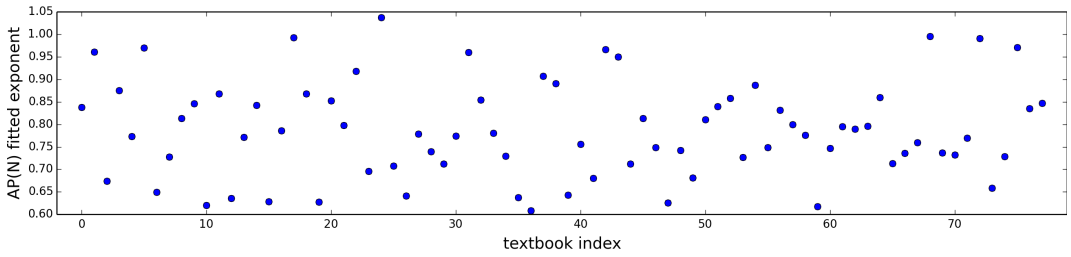


Fig. 5.18 **Power of expansion of $|AP(N(t))|$.** For each textbook, the expansion of the Adjacent possible is fitted as a power function of the total number of tags introduced $AP(N) = aN^{\beta}$. The resulting fitted exponents are reported in the figure, together with the corresponding deviations (even if not visible at this scale).

textbook with id 24 – which corresponds to *Fluid Mechanics* – shows a superlinear expansion of the AP with the number of different tags introduced. In this case, being the exponent slightly over 1, on average, every unit introduced contributes to the almost constant expansion of the adjacent possible space.

Besides the (typically) sublinear trend in the expansion of the adjacent possible as novel units are introduces, its rate of expansion is far from being steady. Of course this is due to the fact that different nodes have a broad different neighbourhood sizes, since the out-degree distribution in Wikipedia is highly heterogeneous.

To obtain a deeper insight into the local exploration of Wikipedia with respect to each textbook content, the above mentioned *adjacent future* observable could be more significant. It is defined as follows. On the Wikipedia graph $G(N,E)$, the subsets of tags belonging to each textbook i is considered, TB_i. According to the definition eq. 5.9 of the actualized set of tags at each time t, it follows that $TB_i \equiv \mathscr{A}(t_{max})$, where t_{max} denotes the final timestamp of the corresponding textbook. Thus, the adjacent future is defined as:

$$AF(t) := \{v \in TB_i \setminus \mathscr{A}(t) : \exists E(w,v) \in G(N,E) \text{ for some } w \in \mathscr{A}(t)\} \qquad (5.11)$$

i.e., at each time it corresponds to the set of tags which will be actualized in the textbook and that are reached by at least one edge from an already actualized tag. Here, the global graph of Wikipedia still serves to define directionality of dependences between nodes.

From the above definition it follows that, contrary to the adjacent possible case, the adjacent future is the null set at the beginning and at the end of any textbook timeline. How it evolves during the narration is significant of the author's strategy of exploration (and) of the content to be presented.

In Figure 5.19, for the usual set of selected textbooks plus a novel one, the size of the adjacent future is reported as a function of the number of distinct tags introduced up to time t. Both axes are rescaled with the total number of tags appearing in each textbook, $|TB_i|$. Different texts explore differently the space of tags they want to propose to the reader. In some cases, the large majority of tags is early ready to be learned by the reader, as the introductory material were presented quite soon in the text. This results in the peaks in the adjacent future size, lying in approximately the same region for the different textbooks considered, around 10-20% from the beginning.

From that moment on the adjacent future can only diminishing in size, more or less constantly. That is the case of reported textbooks with id 12 and 26, respectively corresponding to *College Chemistry* and *General College Chemistry*. On the contrary, in some textbooks like the one with id 76 (*Vibration and Sound*), the adjacent future displays successive enlarging and shrinking phases, thus meaning that the area of the corresponding knowledge space becomes available only during the narration across the text.

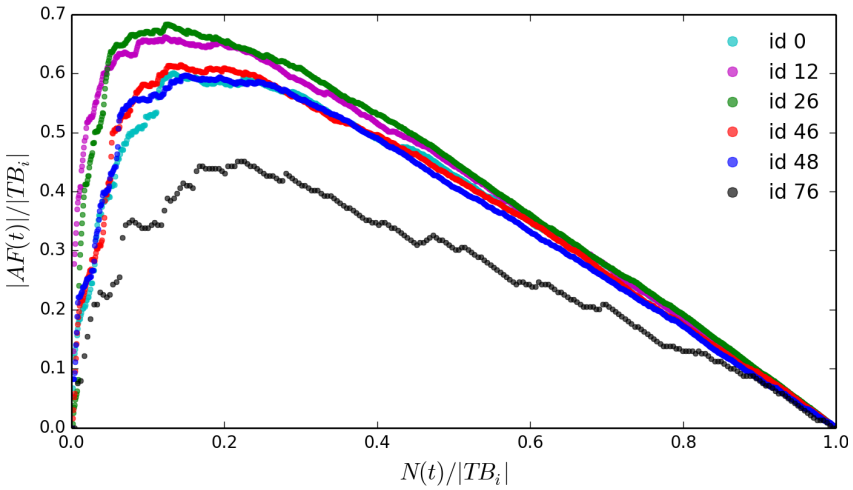


Fig. 5.19 **Adjacent future size evolution as a function of the number of tags introduced.** For the five longest textbooks (ids 0,12,26,46,48) and the textbook with id 76, the fraction of nodes in the adjacent future with respect to the maximum number of tags presentable($|TB_i|$) is reported as a function of the fraction of distinct units introduced to the total $N(t)/|TB_i|$, after averaging over the different stream realizations.

However, the comparison between books belonging to different semantic clusters (with respect to what reported in Figure 5.5) has some severe limitations, partially suggested by the fact that the peaks in the adjacent future size are closer among text semantically closer, i.e., with possibly more tags in common. Indeed, different subsets of tags in the Wikipedia graphs can very differently sampling the underlying distribution, thus revealing the role of the network (sub)structures considered in influencing the present observable.

With this limitation in mind, some further observations are worthy to be made, useful for future analyses. From the differences between the profiles of evolution of the adjacent future across textbooks, both different topics and styles of teaching could be identified. To this end, measures should be introduced to quantify the diverse trends, while textbooks covering similar content could be more deeply analysed in order to bypass the heterogeneity of the underlying network structures. For example, the texts with the most overlapping set of tags introduced could be contrasted.

A second observation regards the profile of modification of the adjacent future along individual texts. Indeed, it would be interesting to recover that spikes and falls in its dynamics correspond to structural sections succession in the original textbooks.

Finally, future works could try to semantically characterize the adjacent learnable, in both the forms introduced in this section, by investigating whether semantics reasoning can reproduce the order by which the adjacent learnable is early covered by exploration.

5.6.2 Cognitive effort

So far, several measures have been introduced to quantified very different aspects of the dynamics of exploration of a *knowledge space*, as suggested by authors in their scientific textbooks. By properly combining the various information provided by the investigations, another key characterization of the textbooks could be recovered, namely its readability. With this, it is referred how hard or easy for the reader is to follow the authors suggested path through the space, i.e., the cognitive effort needed.

Independently of the specificity of the textbook contents, this measure should rather account for the cognitive jumps the reader is forces to do to follow the author narration, and how often and rapidly these jumps are asked.

Given the bunch of quantities introduced in the last sections, one could for example refer to the average semantic distances of consecutive pairs, or on the rate at which the adjacent possible is enlarged but not covered, thus making easier for the reader to deviate from the main path of exploration.

Of course, these are all conjectures, which cannot be validate until independent data from the textbooks readers and users are available. Still, they are here reported as hints for future work.

5.7 Discussion

In this chapter, some preliminary work has been presented on the possible analysis of advanced scientific textbooks in the framework of learning. On the books the investigation proposed has aimed at identifying semantically charged *units*, relevant for representing on an abstract level the content covered in each textbooks. The defined units were chosen to serve as constituting elements of proper knowledge spaces, over which each textbook was supposed to suggest a *learning path* to the reader. Over these paths, analyses were performed to gain insights into their dynamics

of exploration of the knowledge spaces, both from a semantic and a spatial point of view.

It is worth to stress here the originality of the strategies proposed to abstract the texts from the word-level analysis up to the semantically significant level of the proposed *units*. Indeed, to this end, two different approaches were presented and discussed, relying on the subject index of each book or on the external software TAGME [43]. By means of this software, which tags short text fragments with Wikipedia pages, it has been possible to represent each textbook as a dynamical exploration of Wikipedia.

As soon as each textbook can be related to a subset of pages in Wikipedia, their semantic topic representation, as devised in the previous chapter, allows for an aggregate semantic characterization of each book, both in terms of their specificity and interdisciplinarity. Intuitive relations between different books were recovered by computing their semantic distance, thus proving a first proof of the suitability of the topic vectorization procedure.

On the stream of units representing each book narration flow, different observables from quantitative linguistics were considered. The Heap's law was recovered, as well as a very clustered profile of the *units* occurrences along the streams, thus confirming what observed in other linguistic and social systems characterized by an innovation dynamics [108].

Indeed, in every learning path suggested in the textbooks, the reader is supposed to discover novel bits of knowledges. As they are acquired by the reader, her own individual knowledge space enlarges with new nodes or connections. Moreover, other areas of the collective knowledge space, in the present analysis represented by Wikipedia, become *available* for further learning. In this perspective, the Kauffman [68] *adjacent possible* is an *adjacent learnable* by the textbook reader. This approach has been preliminarily tested in this chapter, by proposing two different definitions of the *adjacent learnable*, based on the particular context of Wikipedia. With these, the textbooks exploration dynamics can be quantified, along with the teaching and writing styles of their authors.

In conclusion, while still lacking a definitive quantitative characterization of the texts, innovative tools and approached are explored. Several hints can indeed serve as starting points for future effort in the statistical analysis of the knowledge spaces explorations suggested in any educational course or book.

Chapter 6

Conclusions

Any activity we daily perform while studying, working or just playing requests information, and pushes us to explore *knowledge spaces*. These spaces, far from being only an abstraction, are complex systems where bits of informations interconnect each other, as in the paramount case of Wikipedia [7]. Technological advances have helped us access the knowledge systems, as well as enriching our personal knowledge networks with novel stimuli and associations. Still, the same technologies should provide us innovative tools to exploit the richness of this information space while learning.

To this end, the classical educational schemes should be renewed, in view of the complexity of the information system to which we are exposed. In this perspective, complex systems theory can provide the proper framework where to investigate the complexity features of the knowledge spaces we explore by learning, and the way we usually move in them. This was the scope of the present thesis. The work done was articulated in three major contributions, focused on three different yet complementary aspects of the humans exploration of complex knowledge spaces.

Firstly, the crucial topological properties of the information network to boost the efficiency of a simulated learning exploration were investigated. This was done by devising a class of educational algorithms for scheduling the study practises of a pre-defined collection of items to be-learned, embedded in a network of semantic, linguistic, logical, etc. relationships. While satisfying constraints on the best timing for reviewing and introducing novel material, as suggested by previous research in cognitive science, the algorithms should account also for the possible effects on

learning of the connections between the units. In fact, it was observed in language learners that associations between words can hinder or enhancing the words learning and retrieval. In order to account for this, assumptions had to be done on how to formalize and quantify the possible effects from associations.

After testing the resulting algorithms on different synthetic graphs, it emerged the crucial role of an heterogeneous distribution of connectivity among the nodes. Still, in order the topology to serve for an efficient exploration, a balance between hubs and least connected nodes must be preserved, too specialized items hindering the process. Similarly, the process is impeded by a too clustered topology. Interestingly, these properties are found in the real information networks investigated, namely the free association graph of Human Brain Cloud [49] and some sections of Wikipedia, after removal of the least connected nodes. Their topologies turned out to be almost optimal with respect to some perturbation of their structures. Furthermore, it was found that the order through which the networks are explored as new items are introduced to the learner is essential for taking full advantage of the topology features. Indeed, a random exploration turned out to be ineffective in eliciting the information stored in the graphs.

Of course, the work has some severe limitations. First of all, the assumptions made on how to represent possible effects on learning from associations have not an empirical direct confirmation. Rather, they were inspired by the results obtained in a mere linguistics ambit, which a-priori could not be generalizable to different domains, such as a more articulate information system like Wikipedia. Second, the scheme propose is still too abstract to be immediately comparable with actual learning processes. The same applies to the modelling of the “putative" conceptual networks here considered. Still, the results obtained provide important hints about the role that specific topological structures could have on a class of learning algorithms informed with well-established psychological results. Moreover, the quasi-optimality found of the real information networks considered clearly points to a subtle link between the way in which humans organise their knowledge, i.e., the structure of the knowledge space, and the way in which the information could be retrieved, for instance through a learning path.

In order to further investigate how information is actually retrieved in real information networks, the second contribution discussed focused on the navigation behaviours of Wikipedia readers. Though no real navigation paths are available,

the English Wikipedia Clickstream (EWC) dataset [115] was exploited, gathered by Wikipedia during February 2015 and providing the transition probabilities between pairs of Wikipedia pages. Fed by these transition probabilities, simulated users navigation paths were generated through a memoryless markovian model. To study the simulated walks across pages, a novel abstract *topic* space was defined, whose coordinated are the *main topics* of the Wikipedia classification systems in categories [6]. Thus, still exploiting the category system, a mapping procedure was devised to represent each Wikipedia article as a vector of topics in the mentioned multidimensional topic space. Generalizations of the euclidean distances and the Shannon entropy allowed to characterize the semantic profile of each article, and the semantic distance/similarities between pairs. Such observables, evaluated along the paths, permitted to distinguish between different strategies of the Wikipedia readers, with respect to the source through which the reader enters the encyclopaedia and with respect to other target-oriented information retrieval tasks. Indeed, the real paths generated by players in the online game of Wikispeedia [112] were also considered to contrast the simulated free navigation walks.

Clear patterns emerged in the user's navigation behaviours. Simulated users move differently in the semantic space if their task is goal-oriented – as in Wikispeedia –, if they come from search engines like Google or if they directly enter the encyclopaedia on their pages of interest. Moreover, the longer the walk, the longer the user navigates deeper and deeper levels of specificity, while, independently of the path lengths, the semantic coherence keeps increasing throughout the paths.

Here, the major limitation comes from the simplicity of the assumption made to generate realistic navigation paths. Humans are not markovian while browsing the information networks, and the result found in the semantic topic space confirm previous observations on the need of a semantic representation to fully recover the non-memoryless navigation behaviour of the information-seeker, at least on a aggregate level. Still, a memoryless model from the page probability transitions is the more suitable approach, from a statistical point of view [102], to simulate true navigational paths. Further work could overcome this limitation, if individual real paths of browsing activity would be available for analysis, or more directly inferable from other data. Nevertheless, it is worth to stress the novel approach proposed to represent in a semantic abstract space the Wikipedia articles, thus enriching previous works which did not fully take advantages of the entire category systems of the encyclopaedia.

The proposed schema of semantic abstraction served in the last contribution presented to investigate, semantically, the learning paths suggested by some authors in a bunch of advanced scientific textbooks. Indeed, with the aim at investigating real exploration of knowledge spaces, textbooks were considered as proper ways through field of knowledges, as proposed by their authors to the readers. To the author knowledge, this approach is deeply innovative in the analysis of textbooks, beyond the classical quantitative linguistics aspects.

Indeed, the word level analysis approach was overcome, by identifying semantically charged *units* in each sentences of the texts as nodes in a proper knowledge network, explored by the reader as she comes across the text. The issue of defining the proper semantic units was largely discussed, as well as the two proposed solutions. In particular, the choice of using the articles of Wikipedia as elementary semantic units seemed to be more promising, allowing to again consider Wikipedia as the predefined and collective available knowledge space were to investigate the textbook narrations.

In order to identify in each sentence the more plausible and context related Wikipedia articles, the TAGME [43] platform was used, thus tagging each short sentence with pages from the encyclopaedia. With this, and with the topical representation of each article, the textbook were studied, looking for similarities and field-related statistical features. The preliminary results reported, suggested that Heaps' law [58] holds still at the abstract level of the semantic units, and that different textbooks could correctly distinguished in their contents by their aggregate topic vector representation.

Moreover, first steps are moved in the direction inspired by the Kauffman's *adjacent possible* theory [68]. Could the the texts be investigated as particular realizations of knowledge space exploration? In this sense, different authors can provide different paths across the space, thus diversely enlarging during the narrations the space of the *adjacent learnable* for the reader. Indeed, as she starts the textbook study, the reader is supposed not to know the content presented in the text, at least partially. This can be represented by assuming that the knowledge space is not entirely available for the learner, for example because some prerequisites are missed. As she reads the text, she explores the space with the author, covering the missing preparatories, enlarging her individual known graph and, here the idea of *adjacent learnable*, expanding the space of what she could learn in the future.

All the three proposed lines of research contribute to addressing the stated question on the exploration of knowledge spaces in a large exploratory perspective, mainly proposing and investigating novel tools and approaches. Still, interestingly hints can be drawn from the analyses reported, to be considered both for future research and for future educational applications, improving recommendation systems and classical educational schemes. Only by investigating actual ways of human exploration of information networks, and how semantic relationships witween bits of knowledge are exploited by learners and information seekers, novel tools can be implemented to provide not just content, but true learning paths through it.

References

[1] Duolingo website. www.duolingo.com. Date of access: 09/12/2014.

[2] *Edinburgh Associative Thesaurus* website. http://www.eat.rl.ac.uk/. Date of access: 17/05/2013.

[3] List of main topics used as semantic coordinates in the analysis:. *geography, health, history, humanities, literature, mathematics, nature, people, philosophy, reference works, science, society, technology*.

[4] MediaWiki API. http://en.wikipedia.org/w/api.php.

[5] Wikipedia, dumps. https://dumps.wikimedia.org. "Date of access: 2015-10-22".

[6] Wikipedia, Main Topic Classification. https://en.wikipedia.org/wiki/Category:Main_topic_classifications. Date of access: 2015-10-22.

[7] Wikipedia, the free encyclopedia. http://en.wikipedia.org/. Date of access: 2015-10-22, 2013-11-22.

[8] Wikispeedia. http://cs.mcgill.ca/~rwest/wikispeedia/.

[9] R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 74(47), 2002.

[10] Réka Albert, Hawoong Jeong, and Albert Laszlo Barabási. Diameter of the World-Wide Web. *Nature*, 401(6749):130–131, 1999.

[11] Eduardo G. Altmann and Martin Gerlach. Statistical laws in linguistics. *arXiv*, page 31, 2015.

[12] H.P. Bahrick and L.K. Hall. The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *J. Mem. Lang.*, 52:566–577, 2005.

[13] David A Balota, Janet M Duchek, and Jessica M Logan. Is Expanded Retrieval Practice a Superior Form of Spaced Retrieval? A Critical Review of the Extant Literature. In *The foundations of remembering: Essays in honor of Henry L. Roediger, III*, pages 83–105. Psychology Press, New York, 2007.

[14] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(October):509–512, 1999.

[15] A. Baronchelli, R. Ferrer-i-Cancho, R. Pastor-Satorras, N. Chater, and M. H. Christiansen. Networks in cognitive science. *Trends Cogn. Sci.*, 17(7):348–360, 2013.

[16] M. Barthélemy, A. Barrat, R. Pastor-Satorras, and A. Vespignani. Characterization and modeling of weighted networks. *Physica A*, 346:34–43, 2005.

[17] N. Beckage, L. Smith, and T. Hills. Small worlds and semantic network growth in typical and late talkers. *Plos One*, 6(5):e19348, January 2011.

[18] Nicole M Beckage and Eliana Colunga. Language networks as models of cognition: Understanding cognition through language. In *Towards a Theoretical Framework for Analyzing Complex Linguistic Networks*, pages 3–28. Springer, 2016.

[19] Steven Bird, Edward Loper, and Ewan Klein. Natural language processing with Python, 2009.

[20] M. Boguñá, R. Pastor-Satorras, and A. Vespignani. Cut-offs and finite size effects in scale-free networks. *Eur. Phys. B*, 38(2):205–209, March 2004.

[21] J. Borge-Holthoefer and A. Arenas. Navigating word association norms to extract semantic information. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society, Amsterdam, The Netherlands*, 2009.

[22] J. Borge-Holthoefer and A. Arenas. Semantic networks: structure and dynamics. *Entropy*, 12(5):1264–1302, 2010.

[23] S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Comput. Networks ISDN*, 30(1-7):107–117, April 1998.

[24] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the Web. *Computer Networks*, 33(1):309–320, 2000.

[25] M. Buchanan. *The social atom*. Bloomsbury, New York, NY, USA, 2007.

[26] Luciana Buriol, Carlos Castillo, Debora Donato, Stefano Leonardi, and Stefano Millozzi. Temporal Analysis of the Wikigraph. *2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)*, pages 45–51, dec 2006.

[27] A. Capocci, V. Servedio, F. Colaiori, L. Buriol, D. Donato, S. Leonardi, and G. Caldarelli. Preferential attachment in the growth of social networks: The internet encyclopedia Wikipedia. *Phys. Rev. E*, 74(3):036116, September 2006.

[28] M. Catanzaro and R. Pastor-Satorras. Generation of uncorrelated random scale-free networks. *Phys. Rev. E*, 71(027103):4, 2005.

[29] N. J. Cepeda, H. Pashler, E. Vul, J. T. Wixted, and D. Rohrer. Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychol. Bull.*, 132(3):354–80, May 2006.

[30] Clemesha, Alex. The Wiki Game. http://thewikigame.com/.

[31] A. M. Collins and E. F. Loftus. A spreading-activation theory of semantic processing. *Psychol. Rev.*, 82(6):407, 1975.

[32] A. M. Collins and M. R. Quillian. Facilitating retrieval from semantic memory: The effect of repeating part of an inference. *Acta Psychol.*, 33:304–314, 1970.

[33] Álvaro Corral, Gemma Boleda, and Ramon Ferrer-i-Cancho. Zipf's Law for Word Frequencies: Word Forms versus Lemmas in Long Texts. *Plos One*, 10(7), 2015.

[34] R. G. Crowder. *Principles of learning and memory*. Lawrence Erlbaum Associates, 1976.

[35] Alexander Dallmann, Thomas Niebler, Florian Lemmerich, and Andreas Hotho. Extracting semantics from random walks on wikipedia: Comparing learning and counting methods. In *Tenth International AAAI Conference on Web and Social Media*, 2016.

[36] F. N. Dempster. Spacing Effects and Their Implications for Theory and Practice. *Educ. Psychol. Rev.*, 1(4):309–330, 1989.

[37] Dimitar Dimitrov, Philipp Singer, Florian Lemmerich, and Markus Strohmaier. What Makes a Link Successful on Wikipedia? *Arxiv preprint arXiv:1611.02508*, 2016.

[38] Debora Donato, Stefano Millozzi, S. Leonardi, P. Tsaparas, Universita Roma, La Sapienza, Universita Roma, La Sapienza, Stefano Millozzi, Universita Roma, and La Sapienza. Mining the Inner Structure of the Web Graph. *Proceedings of Eighth International Workshop on the Web and Databases (WebDB 2005)*, 001907(WebDB):152 – 157, 2005.

[39] H. Ebbinghaus. *Memory: a contribution to experimental psychology*. New York City, Teachers college, Columbia university, 1885. trans. H. A. Ruger and C. E. Bussenius, Teachers College at Columbia University, 1913.

[40] D. Elmes. Anki website. http://ankisrs.net/. Date of access: 09/12/2014.

[41] P. Erdös and A. Rényi. On random graphs I. *Publ. Math-Debrecen*, 6:290–297, 1959.

[42] P. Erdös and A. Rényi. On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5:17–61, 1960.

[43] Paolo Ferragina and Ugo Scaiella. TAGME: One-the-fly Annotation of Short Text Fragmetns (by Wikiepdia Entities). *Proceedings of the 19th ACM international conference on Information and knowledge management - CIKM '10*, pages 1625–1628, 2010.

[44] Paolo Ferragina and Ugo Scaiella. Fast and accurate annotation of short texts with wikipedia pages. *IEEE Software*, 29(undefined):70–75, 2011.

[45] R. Ferrer-i-Cancho and R. V. Solé. The small world of human language. *Proc. R. Soc. B*, 268(1482):2261–5, December 2001.

[46] Ramon Ferrer-i-Cancho. The variation of Zipf's law in human language. *European Physical Journal B*, 44:249–257, 2005.

[47] Ramon Ferrer-i-Cancho and Ricard V Solé. Two Regimes in the Frequency of Words and the Origins of Complex Lexicons: Zipf's Law Revisited. *Journal of Quantitative Linguistics*, 8(3):165–173, dec 2001.

[48] Karin Foerde, Barbara J. Knowlton, and Russell A. Poldrack. Modulation of competing memory systems by distraction. *Proceedings of the National Academy of Sciences*, 103(31):11778–11783, 2006.

[49] K. Gabler. *Human Brain Cloud* homepage. http://www.humanbraincloud.com/. Date of access: 09/12/2014.

[50] Florian Geigl, Daniel Lamprecht, Rainer Hofmann-Wellenhof, Simon Walk, Markus Strohmaier, and Denis Helic. Random surfers on a web encyclopedia. In *Proceedings of the 15th International Conference on Knowledge Technologies and Data-driven Business - i-KNOW '15*, pages 1–8, New York, New York, USA, 2015. ACM Press.

[51] Martin Gerlach and Eduardo G Altmann. Stochastic model for the vocabulary growth in natural languages. *Phys. Rev. X*, 3(2):21006, 2013.

[52] Martin Gerlach and Eduardo G Altmann. Scaling laws and fluctuations in the statistics of word frequencies. *New Journal of Physics*, 16(11):113010, 2014.

[53] Bruno Gonçalves and José J. Ramasco. Human dynamics revealed through Web analytics. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 78(2), 2008.

[54] P Gravino, B Monechi, VDP Servedio, F Tria, and V Loreto. Crossing the horizon: exploring the adjacent possible in a cultural system. In *Proceedings of the Seventh International Conference on Computational Creativity*, 2016.

[55] P. Gravino, V. D. P. Servedio, A. Barrat, and V. Loreto. Complex structures and semantics in free word association. *Adv. Complex Syst.*, 15(3 & 4):1250054, 2012.

[56] T. L. Griffiths, M. Steyvers, and A. Firl. Google and the mind predicting fluency with pagerank. *Psychol. Sci.*, 18(12):1069–1076, 2007.

[57] Jacek Gwizdka. Distribution of cognitive load in web search. *J. Am. Soc. Inf. Sci. Technol.*, 61(11):2167–2187, November 2010.

[58] H. S. Heaps. *Information Retrieval: Computational and Theoretical Aspects*. Academic Press, Inc., Orlando, FL, USA, 1978.

[59] T. T. Hills, J. Maouene, B. Riordan, and L. B. Smith. The associative structure of language: contextual diversity in early word learning. *J. Mem. Lang.*, 63(3):259–273, October 2010.

[60] T. T. Hills, M. Maouene, J. Maouene, A. Sheya, and L. Smith. Longitudinal analysis of early semantic networks: preferential attachment or preferential acquisition? *Psychol. Sci.*, 20(6):729–39, June 2009.

[61] D. L. Hintzman. Theoretical implications of the spacing effect. In R. L. Solso, editor, *Theories in cognitive psychology: The Loyola Symposium.* Lawrence Erlbaum, Oxford, 1974.

[62] Todd Holloway, Miran Bozicevic, and Katy Börner. Analyzing and visualizing the semantic coverage of wikipedia and its authors: Research articles. *Complex.*, 12(3):30–40, January 2007.

[63] Petter Holme and BJ Kim. Growing scale-free networks with tunable clustering. *Phys. Rev. E*, (2):2–5, 2002.

[64] Bernardo A Huberman, Peter L T Pirolli, James E Pitkow, and Rajan M Lukose. Strong Regularities in World Wide Web Surfing. *Science*, 280(5360):95–97, 1998.

[65] Eric Jones, Travis Oliphant, Pearu Peterson, et al. SciPy: Open source scientific tools for Python. http://www.scipy.org/, 2001–.

[66] M.A. Just, T.A. Keller, and J. Cynkar. A decrease in brain activation associated with driving when listening to someone speak. *Brain Res.*, 1205:70–80, 2008.

[67] Márton Karsai, Kimmo Kaski, Albert-László Barabási, and János Kertész. Universal features of correlated bursty behaviour. *Scientific Reports*, 2(397), 2012.

[68] Stuart A. Kauffman. Investigations. Working Papers 96-08-072, Santa Fe Institute, August 1996.

[69] G. R. Kiss, C. Armstrong, R. Milroy, and J. Piper. An associative thesaurus of English and its computer analysis. In A. J. Aitkin, R. W. Bailey, and N. Hamilton-Smith, editors, *The computer and literary studies*. University Press, Edinburgh, 1973.

[70] Aniket Kittur, Ed H Chi, and Bongwon Suh. What's in Wikipedia ? Mapping Topics and Conflict Using Socially Annotated Category Structure. *Chi*, pages 1509–1512, 2009.

[71] J Kleinberg, R Kumar, P Raghavan, S Rajagopalan, and A Tomkins. The web as a graph: measurement, models and methods. *Proceedings of the International Conference on Combinatorics and Computing*, pages 1–18, 1999.

[72] Daniel Lamprecht, Dimitar Dimitrov, Denis Helic, and Markus Strohmaier. Evaluating and Improving Navigability of Wikipedia: A Comparative Study of Eight Language Editions. In *Sigcomm 2015*, pages 421–434, 2015.

[73] Daniel Lamprecht, Kristina Lerman, Denis Helic, and Markus Strohmaier. How the structure of Wikipedia articles influences user navigation. *New Rev. Hypermedia Multimed.*, 0(0):1–22, 2016.

[74] Linyuan Lü, Zi-Ke Zhang, and Tao Zhou. Zipf's law leads to Heaps' law: Analyzing their relation in finite-size systems. *Plos One*, 5(12):e14139, 2010.

[75] Matteo Marsili, Iacopo Mastromatteo, and Yasser Roudi. On sampling and modeling complex systems. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(09):P09003, 2013.

[76] Pierpaolo Mastroianni, Bernardo Monechi, Vito D P Servedio, Carlo Liberto, Gaetano Valenti, and Vittorio Loreto. Individual mobility patterns in urban environment. In *Proceedings of the 1st International Conference on Complex Information Systems (COMPLEXIS 2016)*, pages 81–88, 2016.

[77] Mark Meiss, John Duncan, Bruno Gonçalves, J.J. José J Ramasco, and Filippo Menczer. What's in a session: tracking individual behavior on the web. *Ht*, pages 173–182, 2009.

[78] Mark R. Meiss, Filippo Menczer, Santo Fortunato, Alessandro Flammini, and Alessandro Vespignani. Ranking web sites with real user traffic. *Proceedings of the international conference on Web search and web data mining - WSDM '08*, page 65, 2008.

[79] M. Molloy and B. Reed. A critical point for random graphs with a given degree sequence. *Random Struct. Algor.*, 6(2/3):161–179, 1995.

[80] M. Molloy and B. Reed. The size of the giant component of a random graph with a given degree sequence. *Comb. Probab. Comput.*, 7(03):295–305, 1998.

[81] Bernardo Monechi, Pietro Gravino, Vito DP Servedio, Francesca Tria, and Vittorio Loreto. Significance and popularity in music production paper. In prep., 2016.

[82] Elham Najafi and Amir H. Darooneh. The fractal patterns of words in a text: A method for automatic keyword extraction. *Plos One*, 10(6), 2015.

[83] R. Navarro-Prieto, M. Scaife, and Y. Rogers. Cognitive strategies in web searching. In *Proceedings of the 5th Conference on Human Factors & the Web*, 1999.

[84] M. E. J. Newman. Mixing patterns in networks. *Physical Review E*, 67(2):026126, 2003.

[85] M. E. J. Newman. The structure and function of complex networks. *Sirev*, 45(167), 2003.

[86] M. E. J. Newman. Analysis of weighted networks. *Phys. Rev. E*, 70(5):056131, November 2004.

[87] Thomas Niebler, Daniel Schlör, Martin Becker, and Andreas Hotho. Extracting Semantics from Unconstrained Navigation on Wikipedia. *KI - Künstliche Intelligenz*, (December), 2015.

[88] T. P. Novikoff, J. M. Kleinberg, and S. H. Strogatz. Education of a model student. *PNAS*, 109(6):1868–73, 2012.

[89] B. Obama. Remarks by the President at the National Academy of Sciences Annual Meeting. http://www.whitehouse.gov/the_press_office/Remarks-by-the-President-at-the-National-Academy-of-Sciences-Annual-Meeting/, April 2009.

[90] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: bringing order to the web. 1999.

[91] R. Pastor-Satorras, A. Vázquez, and A. Vespignani. Dynamical and correlation properties of the internet. *Phys. Rev. Lett.*, 87(25):258701, November 2001.

[92] P. I. Pavlik and J. R. Anderson. Using a model to compute the optimal schedule of practice. *J. Exp. Psychol.-Appl.*, 14(2):101–17, June 2008.

[93] Alexander M Petersen, Joel N Tenenbaum, Shlomo Havlin, H Eugene Stanley, and Matjaž Perc. Languages cool as they expand: Allometric scaling and the decreasing need for new words. *Scientific reports*, 2, 2012.

[94] P. Pimsleur. A memory schedule. *Mod. Lang. J.*, 51(2):73–75, 1967.

[95] Jacob Ratkiewicz, Filippo Menczer, Santo Fortunato, Alessandro Flammini, and Alessandro Vespignani. Traffic in Social Media II: Modeling Bursty Popularity. *2010 IEEE Second International Conference on Social Computing*, pages 393–400, aug 2010.

[96] GC Rodi, V Loreto, VDP Servedio, and F Tria. Optimal learning paths in information networks. *Scientific Reports*, 10286, 2015.

[97] Giovanna Chiara Rodi, Vittorio Loreto, and Francesca Tria. Search strategies of wikipedia readers. *PLOS ONE*, 12(2):1–15, 02 2017.

[98] Malte Schwarzer, Moritz Schubotz, Norman Meuschke, and Corinna Breitinger. Evaluating Link-based Recommendations for Wikipedia. *Proc. 16th ACM/IEEE Jt. Conf. Digit. Libr.*, pages 191–200, 2016.

[99] Tom A Schweizer, Karen Kan, Yuwen Hung, Fred Tam, Gary Naglie, and Simon Graham. Brain activity during driving with distraction: an immersive fmri study. *Frontiers in Human Neuroscience*, 7(53), 2013.

[100] M Angeles Serrano, Alessandro Flammini, and Filippo Menczer. Modeling statistical properties of written text. *Plos One*, 4(4):e5372, 2009.

[101] Choi Seung-Seok, Cha Sung-Hyuk, and Charles C Tappert. A Survey of Binary Similarity and Distance Measures. *Journal of Systemics, Cybernetics & Informatics*, 8(1):43–48, 2010.

[102] Philipp Singer, Denis Helic, Behnam Taraghi, and Markus Strohmaier. Detecting memory and structure in human navigation patterns using Markov chain models of varying order. *Plos One*, 9(7), 2014.

[103] Rv Solé, B Corominas-Murtra, Sergi Valverde, and Luc Steels. Language networks: Their structure, function, and evolution. *Complexity*, (22):1–9, 2010.

[104] Massimo Stella, Nicole M Beckage, and Markus Brede. Multiplex lexical networks reveal patterns in early word acquisition in children. *arXiv preprint arXiv:1609.03207*, 2016.

[105] M. Steyvers and J. B. Tenenbaum. The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cognitive sci.*, 29(1):41–78, January 2005.

[106] Krzysztof Suchecki, Alkim Almila Akdag Salah, Cheng Gao, and Andrea Scharnorst. Evolution of Wikipedia's Category Structure. *Advances in Complex Systems*, 15(supp01):1250068, jun 2012.

[107] T. Tinkham. The effects of semantic and thematic clustering on the learning of second language vocabulary. *Second Lang. Res.*, 13(2):138–163, April 1997.

[108] Francesca Tria, Vittorio Loreto, Vito DP Servedio, and Steven H. Strogatz. The dynamics of correlated novelties. *Nature Scientific Reports*, 4(5890), 2014.

[109] Michael S Vitevitch. What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech Language and Hearing Research*, 51(April):408–422, 2008.

[110] D J Watts and S H Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–2, jun 1998.

[111] Robert West and Jure Leskovec. Human wayfinding in information networks. *Proceedings of the 21st international conference on World Wide Web - WWW '12*, page 619, 2012.

[112] Robert West, Joelle Pineau, and Doina Precup. Wikispeedia: An Online Game for Inferring Semantic Distances between Concepts. *IJCAI*, pages 1598–1603, 2009.

[113] Jake Ryland Williams, James P Bagrow, Christopher M Danforth, and Peter Sheridan Dodds. Text mixing shapes the anatomy of rank-frequency distributions. *Physical Review E*, 91(5):052811, 2015.

[114] P. A. Woźniak and E J Gorzelańczyk. Optimization of repetition spacing in the practice of learning. *Acta Neurobiol. Exp.*, 54(1):59–62, January 1994.

[115] Ellery Wulczyn and Dario Taraborelli. Wikipedia Clickstream. 10 2016.

[116] Damián H Zanette. Statistical patterns in written language. *arXiv preprint arXiv:1412.3336*, 2014.

[117] GK Zipf. Human behavior and the principle of least effort. *Ed: Addison-Weslay, Reading, MA*, pages 268–270, 1949.

[118] V. Zlatić, M. Božičević, H. Štefančić, and M. Domazet. Wikipedias: Collaborative web-based encyclopedias as complex networks. *Phys. Rev. E*, 74:016115, Jul 2006.

Appendix A

Network theory: terminology

This chapter reviews some concepts and definitions from network theory, used throughout the thesis. For a more comprehensive survey on complex networks and their properties, please refer to [9, 85, 16, 86].

A.1 Notation and observables

A network is a graph contituted by vertices (nodes) and edges (links) connecting them. Using a notation borrowed from graph theory, it is typically denoted by $G(V,E)$, where V and E are the set of nodes and edges respectively. The edges can carry different types of information, typically directionality and weight. Depending on this, networks can be (un)directed and (un)weighted. If not differently specified, any edge e_{ij} is a direct link from node i to j. Its weight is denoted by w_{ij}. In case of unweighted network, it holds: $w_{ij} = 1 \, \forall i, j \in V$.

Connectivity

For each node i, its degree deg_i and strength s_i hold information about its connectivity. In case of undirected graph, they are defined as the number of links connected to that node and the sum of their weights, respectively. If the network is directed, degree (and thus also strength) can be specialized to account only for the links incoming or outgoing in/from each node.

By introducing the adjacency matrix $\mathbf{A}$, whose (i,j)-th element a_{ij} is 1 if nodes i and j are connected, 0 otherwise, the *in-degree* and *out-degree* are defined:

$$deg_i^{in} = \sum_{j=1}^{N} a_{ji} \tag{A.1}$$

$$deg_i^{out} = \sum_{j=1}^{N} a_{ij} \tag{A.2}$$

with N number of nodes, i.e. network order: $N = |V|$. The strengths are similarly defined, after multiplying the adjancency matrix elements a_{ij} with the corresponing edge weights w_{ij}. It follows that, in unweighted graphs, the two quantities coincide.

The neighbourhood of node i, is the set of vertices direclty connected to node i. In case of directed networks:

$$\mathcal{N}_i^{in} = \{j \in V : \exists\, e_{ji} \in E\} \tag{A.3}$$

$$\mathcal{N}_i^{out} = \{j \in V : \exists\, e_{ij} \in E\}. \tag{A.4}$$

Components

Given the connectivity profiles of all nodes, it is possible to identify different regions in the newtroks, depending on the reachability of their nodes. A graph $G = (V,E)$ is said to be:

- *strongly connected* if for every pair of node a path connecting them exists;
- *connected* if for any pair of nodes (i,j), either i is reachable from j or j is reachable from i;
- *weakly connected* if its undirected version is connected.

In case of undirected graph, the definitions coincide. If one of the previous cases is verified only on a subgraph G' of the original graph, it is said to be a strongly/weakly connected component of G.

Between any two node a distance can de defined. Typically, the *shortest path* is considered, i.e., the length of the shortest path among the ones connecting the nodes. If two nodes are not reachabled or they belong to different components, their distance is set to infinity.

K-cores decomposition

The subgraph $G'(V',E')$ is said to be the k-core of G if it is the maximal connected subgraph whose nodes have degree equal or greater than k. It can be extracted from a $(k-1)$-core by iteratively filtering out nodes with degree smaller than k.

Clustering coefficient

The *clustering coefficient* $c.c.$ allows to quantify the network transitivity, i.e. the probability that two nodes sharing a neighbour are connected. Two definitions of the coefficient can be considered [85]. The first one refers to the fraction of triangles in the network up to the number of the possible triples. The second one, hereafter considered, consists in evaluating for each node i its local clustering $c.c._i$

$$c.c._i = \frac{\text{\# triangles connected to node } i}{\text{\# triples centered on node } i} \tag{A.5}$$

and then averaging it over the N graph nodes:

$$c.c. = \frac{1}{N}\sum_i c.c._i. \tag{A.6}$$

Assortative mixing

With *assortative mixing* the tendency of nodes with similar properties to be connected is usually indicated. When the similar property is the degree, the (dis)assortative nature of a graph implies the presence of positive (negative) degree-degree correlations.

Different tools are available to investigate the graph assortative nature. For example, the conditional probability that a node with degree deg' is linked to a node with degree deg, $P(deg'|deg)$, is usually considered. Numerically, it is quantified by the nearest neighbour average degree of nodes with degree deg:

$$\langle deg_{nn}(deg)\rangle = \sum_{deg'} deg' P(deg'|deg). \tag{A.7}$$

For uncorrelated networks $P(deg'|deg) \propto deg' P(deg')$ and thus $\langle deg_{nn}(deg)\rangle$ becomes independent of the degree deg. On the contrary, an increasing (decreasing)

dependence of $\langle deg_{nn}(deg)\rangle$ on deg is a signature of positive (negative) degree-degree correlations, i.e. of assortative (disassortative) mixing in the network [91].

In contrast, the correlations can be quantified through diverse *assortative indexes* (a) [85]. In the thesis, the following scalar assortativity is considered, for undirected networks:

$$a = \frac{1}{\sigma_q^2}\sum_{jk} jk(e_{jk} - q_j q_k), \tag{A.8}$$

where j,k runs on the possible node degrees, q_j is the probability distribution that an edge leads to a node with degree $(j+1)$, e_{jk} is the joint probability distribution that an edge connects two nodes with degrees $(j+1)$ and $(k+1)$, and σ_q is the variance of the distribution q_k, so that $-1 \leq r \leq +1$.

A.2 Generative models

In this section, the generative models reffered to in the present manuscript are described.

A.2.1 Random graph

The random graph model proposed by Erdös and Rényi [41, 42] has been considered and graphs of the ensamble $G_{N,m}$ have been generated in the following way. In a set of N vertices, m links are randomly placed, where m is a fraction of the possible node pairs, $m = p \cdot \frac{1}{2}N(N-1)$, depending only on the connection probability p. In the infinite size limit, this approach leads to a Poisson-like degree distribution with mean degree $\langle deg\rangle = p(N-1)$:

$$P(deg) = \binom{N}{deg} p^{deg}(1-p)^{N-deg} \simeq \frac{\langle deg\rangle^{deg}\exp^{-\langle deg\rangle}}{deg!}. \tag{A.9}$$

Graphs so generated have neither assortativite nor disassortative mixing.

A.2.2 Scale free graphs

Barabási-Albert model

As a classic example of a scale-free network, the model proposed by Barabási and Albert [14] has been considered. Starting with a full-connected set of m_0 nodes, a network of N nodes is built, by adding one-by-one $N - m_0$ nodes each of them with m ($m < m_0$) edges to be linked in the network. Each edge is attached to another existing vertex according to the *preferential attachment* principle, i.e. with probability proportional to its degree. This growth rule yields to a small-world network with a power-law degree distribution $P(deg) \sim deg^{-\gamma}$, the exponent being $\gamma = 3$. The resulting clustering coefficient is higher than in a random graph with of the same order and size.

Holme-Kim model

The generation model proposed by Home and Kim[63] allows to create graphs with properties very similar to the standard scale-free graphs, i.e., power-like degree distribution and small world effects (small average geodesic distances), but with a modifiable clustering coefficient. Indeed, the transitivity of the resulting graphs can be tuned by a control parameter. In the generation rules, together with the preferential attachment, a principle of triad formation is considered as follows. Starting from a completely disconnected set of m_0 nodes, at each step a new node u is added with m links. The first link is connected to another existing vertex according to the *preferential attachment* principle (PA), as in BA model, say v. Then, for each of the $m-1$ edge to be attached, with probability P_t a *triad formation* step is performed, i.e., a neighbor of the last node connected (v here) is chosen to receive the edge, thus creating a triad, if possible. Otherwise, with probability $1 - P_t$ a novel node is chosen with a PA step. The probability parameter P_t allows to tune the final clustering coefficient of the network.

Uncorrelated configuration model

In order to explore the behaviour of scale-free networks with different power-law distribution exponents, the *Uncorrelated Configuration Model* (UCM) proposed by Catanzaro et al. [28] has been considered. Through this model, which is based on the

more general *Configuration Model* (CM) [79, 80], scale-free networks are generated with no degree-degree correlations.

As in the CM, each node n in a set of N is assigned with a random degree deg_n extracted from a known distribution $P(deg)$, where $m_0 \leq deg_n \leq N$ and the sum $\sum_n deg_n$ must be even. The nodes are then randomly linked respecting their preassigned degrees. If the distribution is no bounded, and the fluctuations $\langle deg^2 \rangle$ diverge in the infinite network order limit, as in the case of a scale-free distribution $P(deg) \sim deg^{-\gamma}$ with $\gamma \leq 3$, the network resulting by applying the procedure described is uncorrelated. But, if another constraint is made, i.e. if self-loops and multiple connections are forbidden [20], the resulting network presents a disassortative mixing for high-degree nodes. In order to avoid these correlations, in UCM, a structural cut-off on the maximum degree is imposed which scales with the network order as $N^{1/2}$. With this new constraint on the assignable degrees $m_0 \leq deg_n \leq N^{1/2}$, uncorrelated scale-free networks with neither self-loops nor multiple connections are generated.