Self-emergence of knowledge trees: Extraction of the Wikipedia hierarchies

Lev Muchnik,[1] Royi Itzhack,[2] Sorin Solomon,[3,4] and Yoram Louzoun[2,*]
[1]*Physics Department, Bar Ilan University, Ramat Gan, Israel 52900*
[2]*Math Department, Bar Ilan University, Ramat Gan, Israel, 52900*
[3]*Racah Institute of Physics, Hebrew University, Jerusalem, Israel*
[4]*ISI, Torino I-10133, Italy*

(Received 25 July 2006; revised manuscript received 15 January 2007; published 13 July 2007)

The rapid accumulation of knowledge and the recent emergence of new dynamic and practically unmoderated information repositories have rendered the classical concept of the hierarchal knowledge structure irrelevant and impossible to impose manually. This led to modern methods of data location, such as browsing or searching, which conceal the underlying information structure. We here propose methods designed to automatically construct a hierarchy from a network of related terms. We apply these methods to Wikipedia and compare the hierarchy obtained from the article network to the complementary acyclic category layer of the Wikipedia and show an excellent fit. We verify our methods in two networks with no *a priori* hierarchy (the *E. Coli* genetic regulatory network and the *C. Elegans* neural network) and a network of function libraries of modern computer operating systems that are intrinsically hierarchical and reproduce a known functional order.

DOI: 10.1103/PhysRevE.76.016106 PACS number(s): 89.75.Fb, 89.75.Hc

I. INTRODUCTION

Throughout history, beginning with Aristotle's ontology [1], knowledge was repeatedly systematized into well-defined structural hierarchies. One of the outstanding landmarks of this process is the modern encyclopedia, which was developed by a group of intellectuals in the 18th century who called themselves *Encyclopedists*. The *Encyclopedie* [2] arranged all human knowledge at the time in a hierarchical structure, allowing sensible navigation and orientation within that knowledge. Essentially, the hierarchical structure enabled users to access knowledge with great ease. On account of these attributes, the *Encyclopédie* can be considered one of the great catalysts of modern science [3].

The rate at which knowledge is accumulated has changed since the creation of the *Encyclopédie*. Today, knowledge is accumulated in an exponential manner [4,5]. As such, knowledge can no longer be summarized by a small group of researchers into a well-defined hierarchy. For example, *Encyclopédie* contained 71 818 entries compared with over one million in the current English Wikipedia. Today's rapid knowledge accumulation has led to a change in the way knowledge is arranged, spurring the development of qualitatively different repositories, such as the WWW [6], Wikipedia [7], and PubMed [8]. The construction and use of today's knowledge repositories differ drastically from that of earlier repositories like the *Encyclopédie*. In today's knowledge repositories, the arrangement, collection, and sometimes even the content of the information being contributed by uncoordinated sources is generally not subject to centralized moderation. The carefully designed structure of the previous generation of knowledge repositories is lost in this newer process. Thus, the knowledge hierarchy may be treated as an obsolete relic of the time when the amount of knowledge was limited.

The hierarchical structure of knowledge is not merely an artificial structure built to ease the access to it. It actually represents the more inherent fact that human knowledge is based on generalizations from a set of details into a general rule. These simplifications (which can often oversimplify reality) allow us to reach conclusions that, while not entirely precise, are frequently broad enough to provide insight. Thus, while it is now unfeasible to manually build knowledge hierarchies, the need for them is even more crucial. The solution may come directly from the accumulated data. The uncoordinated deposition of knowledge may actually lead to self-emerging hierarchies, not dictated by a limited group of people sharing a similar perception, but directly by the information. In the present work, we propose algorithms to extract the underlying hierarchical structures concealed within a knowledge repository and automatically classify terms in different positions of the hierarchy. We test these algorithms on the Wikipedia networks and show their precision. We further show that these algorithms can be effectively employed to construct a hierarchy on networks from various domains.

We base our approach on the assumption that the hierarchical position of concepts is not only a function of their content, but also of their context (*France* is a subcategory of *state*, only because there are many others like it). While the content cannot reveal the full context, context may contain information about the content. We here show that the context information is enough to extract an approximate hierarchy, ignoring the content of each term by itself.

We address context by looking at all terms in a knowledge repository as a directed network. This directed network contains acyclic subnetworks. Any acyclic network can be treated as a hierarchy, starting from nodes with no incoming edges toward nodes without outgoing edges. We here propose a few approaches to choose such subnetworks, so that they represent a meaningful hierarchy.

We have chosen to use the Wikipedia [9] in 23 different languages to show that the network structure can be used to reproduce a human-designed hierarchical structure. The Wikipedia structure has recently been subject to extensive

*louzouy@math.biu.ac.il

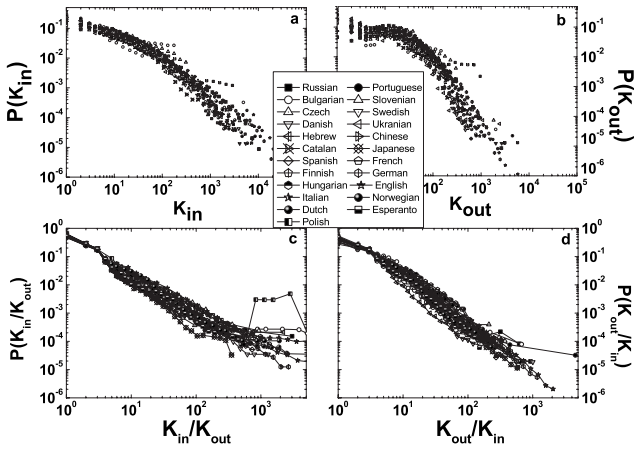

FIG. 1. Incoming (a) and outgoing (b) node degree distribution of Wikipedia-induced networks in different languages. Distribution of incoming divided by outgoing nodes' degree (c) for $k_{\mathrm{in}} > k_{\mathrm{out}}$ and outgoing divided by incoming degree (d) for $k_{\mathrm{out}} > k_{\mathrm{in}}$.

study [10,11]. It is freely edited by the internet community, and just like the WWW is generated by a myriad of individual editors [13 000 contributors with more than five entries were recorded in only 1 month (January 2005)]. Like the WWW [12], it contains the hallmarks of complex emerging systems, such as scale free in-degree distribution [Fig. 1(a)] and fat tail out-degree distribution—although the out-degree distribution shows a much steeper power than the in degree [Fig. 1(b)]. In contrast with the WWW, Wikipedia has a collective goal and an integrated structure. It is built to represent human knowledge in an encyclopedic way. Thus, practically all Wikipedia pages are a part of the human knowledge web.

The Wikipedia also contains a human designed hierarchical concept structure—the Wikipedia categories [13]. This structure is also edited by the Wikipedia editors. However, in contrast with the articles, it is explicitly acyclic, directed from narrow towards more general terms. We have actually checked that the category network contains no cycles (Table III, column 12). The presence of this complementary representation allows us to directly test the validity of algorithms based on the proposed approaches.

Wikipedia exists in over 200 languages and cultures. A comparison between the different languages can differentiate fundamental from contingent properties [11]. In general, only robust network properties can be used to extract a hierarchy. We will thus base the hierarchy measures on properties found to be common to Wikipedia in all languages.

In the following sections, we first reproduce and enlarge the Wikipedia network property measurements. We then define and validate five hierarchy definition algorithms.

II. METHODS

A. Data

Since Wikipedia is intrinsically a public project, standing for free access to the treasury of human knowledge, the Wikimedia Foundation Inc. [9] periodically provides complete snapshots of the 216 Wikipedias in different languages. The most up-to-date snapshot is generously available from Ref. [14], while the data used in the present paper represents the snapshot taken on 22 April 2005. Along with regular encyclopedic articles, each Wikipedia contains a wider range (16) of entry types (e.g., user pages, talks and discussions, categories, etc.). In the current work, we analyze directional networks obtained solely from the article entries. We validate some of our conclusions by comparing them with the explicit, human designated hierarchy of the categories.

B. Translation of Wikipedia to network

Each Wikipedia contains a vast amount of content and information, which was reduced to a sheer skeleton: a set of nodes representing the Wikipedia articles and a collection of directed edges representing the hypertext references between the articles. Each edge points from the node representing the article that quotes another article to the node representing the quoted article. Articles that redirect one another because they correspond to the same entry, or nodes resulting from abbreviation, misspelling and different capitalization (e.g., USA redirects to "United States of America") were merged to a single node (i.e., there is a unique node in our network for USA and for "United States of America"). Nodes that have no incoming edges (about 13% of the nodes) are removed from the analysis.

C. Network analysis toolbox library

We have developed a comprehensive network analysis toolbox library containing graph import and generation tools, graph manipulation routines, and network analysis tools to carry out a large number of network measurements across many applications and disciplines. This MatLab based toolbox provides a coherent, easily expendable environment, which to the best of our knowledge is the most complete graph analysis package available today. At present, it implements over 60 analytical and utility methods for graph analysis and manipulation.

D. Distribution and shortest path length average

The distance $d(i,j)$ between nodes i and j is defined as the number of edges in the shortest (geodesic) path containing i and j. Note that on a directed graph, the distance $d(i,j)$ from i to j is not always equal to the distance $d(j,i)$ between j and i. In fact, both are longer or equal to the naive undirected shortest path, since the undirected path often violates the correct edges direction.

E. Clustering coefficient

Given the set of $k(i)$ neighbors of a node i, one could in principle have $k(i)[k(i)-1]/2$ undirected edges between them. The clustering coefficient $[C_{C(i)}]$ of a node i measures what fraction of this set is actually connected. For a large, completely random graph, $\langle C_{C(i)} \rangle$ is $\langle k \rangle / N$, while for a completely connected graph $C_{C(i)}$ is 1. Together with the shortest path distribution, the clustering coefficient is a measure of

the "small world" property as implemented by Watts and Strogatz [15]. We have measured both undirectional and directional clustering coefficients. The directional clustering coefficient was computed using just incoming edges or just outgoing edges.

F. Betweenness centrality

Betweenness centrality has frequently been used to estimate network resilience [16]. Given all geodesic paths in a network, the betweenness centrality $C_B(v)$ of a node $v \in V$ (V is a set of network nodes) is defined as the fraction of the shortest paths passing through v, $C_B(v)=\sum_{s\neq v\neq t\in V}\frac{\sigma_{st}(v)}{\sigma_{st}}$ where σ_{st} denotes the number of shortest paths from node $s \in V$ to node $t \in V$, and $\sigma_{st}(v)$ is the number of shortest paths from node s to t passing through $v \in V$. The sum is executed over all pairs of network nodes. We have computed the betweenness centrality using the highly efficient Brandes algorithm [17].

G. PageRank

PageRank, as defined by Page *et al.* [18], is designed to compute the rank of each node supposedly corresponding to its importance in the network. PageRank regards directed edges in a network as votes for destination node significance. PageRank recursively defines the rank of a node as the weighted sum of the edges pointing to it, and the weight of an edge as the weight of its origin node divided by the origin node outgoing degree. PageRank computation is initiated by assuming an arbitrary initial rank distribution among the nodes and iteratively computing their new rank until the procedure converges (usually within a very small number of iterations, not exceeding 20).

H. HITS

We have also applied the HITS (hypertext induced topic selection) algorithm [19] to Wikipedia. HITS assigns two qualities to each node: *authority* and *hub*. It is assumed that a node has a high authority if it is pointed out by many valued hubs. A node is a highly ranked hub if it points at nodes with high authority. Similar to the PageRank, HITS iteratively updates the (initially arbitrary) authority and hub values of each node until it converges.

I. Circle enumeration

A directed circle is a closed directed path containing no repetitions (i.e., none of the nodes in the circle appear 2 times). We define a minimal circle as a circle composed of a minimal path from one node to the other and the minimal path back to the original node. Since the same node can appear in many circles, we developed an algorithm that would count each node only once (i.e., if a circle of size r would be counted j times, the circle counter of size r increases only by $1/j$, every time we measure this circle). The precise details of the algorithms will be presented in a future paper.

III. RESULTS

We here use Wikipedia as a test case for extracting a meaningful hierarchy from the network context. Wikipedia is represented here as a directed network in which nodes correspond to articles and edges to references. The large amount of data present in the Wikipedia and the precise definition of the network allow us to overcome the limitations found in previous attempts to construct and study concept structures [20,21]. Such limitations include the dependence on various temporal, individual, and cultural factors, and the imprecise definition of an edge between two concepts.

We performed a comparison of the network analysis of the article network of Wikipedias in 23 languages, which had between 8500 and 940 000 articles at the time of the snapshot (April 2005). Assuming that hierarchy must manifest itself in similar ways in all cultures, we searched for a set of properties which remained invariant in all tested Wikipedias. The network properties of Wikipedia were previously studied by multiple groups [10,11]. Following discrepancies between the different groups results (e.g., the outgoing degree power was measured to be 2.1 Vs 2.6), we have reproduced all measures as well as a few new measures required for the current analysis. We find that the extensive properties vary widely among Wikipedias, while most intensive (properties not scaling with the networks size) measures are consistent among different languages. For example, the number of nodes varies from 8500 to 900 000. The average number of edges per node varies from 8.5 to 29 and is weakly correlated with the network size (Table I). The in degree exhibits a fat tail controlled by a scale-free distribution beyond a minimal value of 100 incoming edges, as is the case in most complex networks [Fig. 1(a)] [12,22–26]. The out degree has a fat tail, which is not perfectly scale free (Fig. 1). The geodesic path length distribution is identical in all tested languages and is short (Fig. 2). The (directed) clustering coefficient varies [27], but is extremely high (0.14–0.26) (Table I). The last two elements combined correspond to the *small world* property of all known conceptual networks [15]. Notice, that the high clustering coefficient is not accompanied by any scaling behavior. The clustering coefficient is rising from a minimal value for nodes with a low degree to a maximal value around a degree of 100 and then a drop to 0 for high degrees (Fig. 3). A special aspect of the Wikipedia is the huge fraction of reciprocal links. In average, if i points to j, there is a 23%–59% (depending on language) probability that j points back to i (Table II).

We have found two measures perfectly scaling between Wikipedias in different languages:

(i) The $k_{\text{in}}(i)/k_{\text{out}}(i)$ distribution, which exhibits a much clearer power law than each separately [Figs. 1(a) and 1(b) vs 1(c) and 1(d)]. This invariant suggests that the ratio between in and out degree is a better indicator of content than the degree itself. Similar ideas have been widely used to rank WWW sites in search engines [19] (where the dual role of hubs and authorities has been exploited).

(ii) The betweenness centrality which is closely related to node's degree (Fig. 5) and scales with network size and average degree (Figs. 5 and 6).

TABLE I. Summary of the basic size-related measurements of reference networks of 23 Wikipedias in different languages (columns 2 and 3). Column 4 lists the total number of unique articles stored in the Wikipedia database we analyzed. Column 5 displays the number of articles having no content and redirecting the reader to the proper page. The net number of stand-alone articles is listed in column 6 and the percentage of the redirections in each language appears in column 7. All other measurements in the paper are performed on graphs from which redirecting nodes have been removed and links pointing to and from them are relinked to the appropriate nodes representing the redirecting one. Total number of links and average node degree are listed in columns 8 and 9. Total number of edits, number of registered users, and average number of edits per registered users are reported in columns 10, 11, and 12 accordingly. Notice that some of the edits in column 10 have been performed by unregistered users. Those were subtracted when the average user activity was computed.

Number	Language	Code	Bulk number of articles	Number of redirections	Net number of nodes	Percent of redirections	Number of links	Average degree	Number of edits	Number of registered users	Average number of edits per user
1	Russian	Ru	21325	1398	19927	6.6	275657	13.8	126133	1159	108.8
2	Bulgarian	bg	29155	1250	27905	4.3	449931	16.1	97361	374	260.3
3	Czech	cs	9105	577	8528	6.3	122680	14.4	43701	400	109.3
4	Danish	da	34329	4617	29712	13.4	298931	10.1	165372	756	218.7
5	Hebrew	he	27494	3045	24449	11.1	380222	15.6	219494	1338	164
6	Catalan	ca	18370	2516	15854	13.7	164455	10.4	64648	385	167.9
7	Spanish	es	63134	5658	57476	9	816734	14.2	451354	3946	114.4
8	Finnish	fi	26894	4067	22827	15.1	375581	16.5	137652	973	141.5
9	Hungarian	hu	10789	1810	8979	16.8	149092	16.6	66547	315	211.3
10	Italian	it	49728	2981	46747	6	941425	20.1	370815	2028	182.8
11	Dutch	nl	93854	13607	80247	14.5	1373475	17.1	800813	3311	241.9
12	Polish	pl	85803	7941	77862	9.3	2233878	28.7	569408	3141	181.3
13	Portuguese	pt	50654	3904	46750	7.7	569909	12.2	223348	1789	124.8
14	Slovenian	sl	14462	773	13689	5.3	358787	26.2	81520	221	368.9
15	Swedish	sv	90523	8607	81916	9.5	1057964	12.9	419868	1621	259
16	Ukranian	uk	19103	556	18547	2.9	156806	8.5	43245	138	313.4
17	Chinese	zh	40939	14744	26195	36	504260	19.3	302644	3264	92.7
18	Japanese	ja	162607	44613	I117994	27.4	2566127	21.7	1496716	5844	256.1
19	French	fr	136818	12148	124670	8.9	2469504	19.8	1409097	7280	193.6
20	German	de	342511	61258	281253	17.9	5777619	20.5	4043185	29075	139.1
21	English	en	993562	98372	895190	9.9	13495590	15.1	9336036	64588	144.5
22	Norwegian	no	30923	3098	27825	10	342787	12.3	123600	852	145.1
23	Esperanto	eo	28249	2571	25678	9.1	320759	12.5	119971	557	215.4

The precise scaling of these measures hints that these measures may be the bearer of content-dependent information. These properties are natural first guesses to be used for hierarchy detection (Fig. 4).

A simple way to obtain a hierarchical structure is to define an appropriate score for each node and two cutoffs. One can then compare the ratio of scores of two neighboring nodes (in the underlying undirected network), and define that a node is higher in hierarchy than its neighbor (independently of the edge direction) if its score is higher and their score ratio is between the lower and upper threshold. Many neighboring nodes have no hierarchical relation. Their score ratio can be too close to one and thus above the upper threshold (e.g., cheese and meat). Their ratio could also be too far from one and thus below the lower threshold (e.g., U.S. and Daleville, Alabama). Note that there is probably no link from U.S. to Daleville, but there is one from Daleville to U.S., so that in the underlying undirected network, they are con-

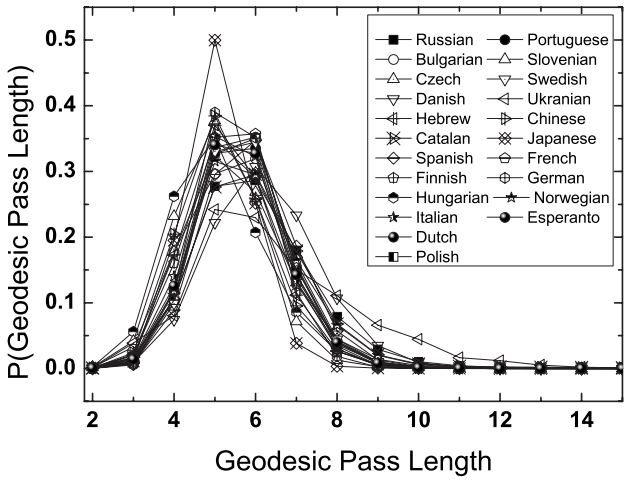

FIG. 2. Shortest pass length distribution of Wikipedia articles network in different languages.

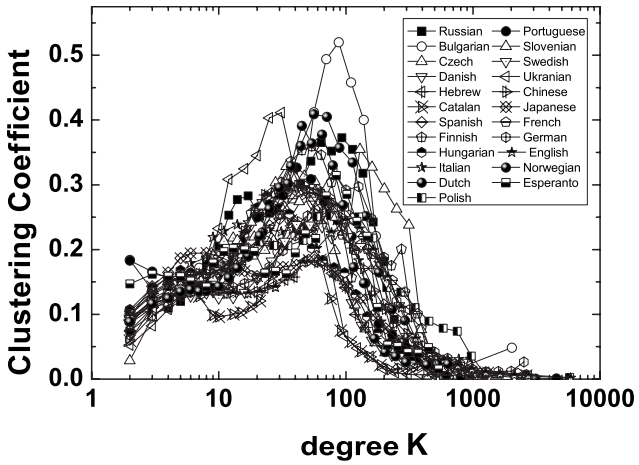

FIG. 3. Average clustering coefficient as a function of nodes' degree for different Wikipedia languages.

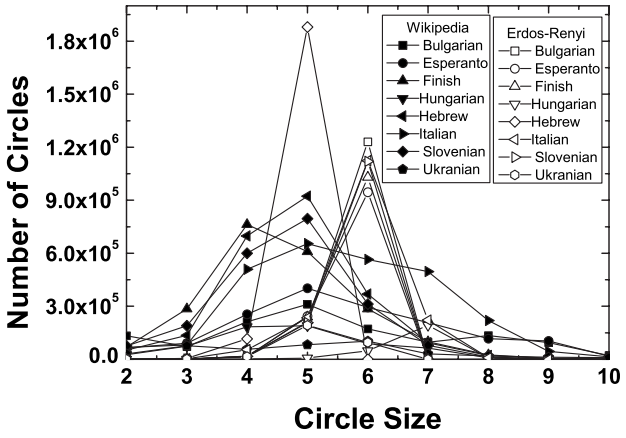

FIG. 4. Circular pass lengths for Wikipedia articles network (filled markers) and Erdos-Reny (ER) networks (hollow markers) of similar sizes and average degree. Circles in ER networks are noticeably longer.

TABLE II. More detailed measurements of the Wikipedia reference networks. Columns 3 and 4 list the α of the fit of the k_{in}/k_{out} distributions (for $k_{in}>k_{out}$, column 3) and of k_{out}/k_{in} distribution ($k_{in}<k_{out}$, column 4) to the power law of the form $P(k)\sim P^{-1-\alpha}$. The appropriate distributions are shown in Figs. 1(a) and 1(b). Columns 5–7 list the average clustering coefficient of each graph as defined in the paper. Column 8 details the percentage of links in the graph that are coupled with links going in the opposite direction. Column 9 (10) shows the percentage of nodes having no incoming (outgoing) links. Columns 11–13 show an estimate of the average shortest pass length when the links are followed in their original direction (column 11), inverse (column 12) and when the direction is ignored.

				Average clustering coefficient						Average shortest distance		
Number	Code	$P(k_{in}/k_{out})$ α	$P(k_{out}/k_{in})$ α	Direct	Inverse	Both	Reciprocity %	$P(k_{in}=0)$ %	$P(k_{out}=0)$ %	Direct	Inverse	Both
1	Ru	1.4	1.3	0.21	0.22	0.24	40	15	1	5.1	5.1	5.1
2	Bg	1.3	0.9	0.24	0.24	0.26	59	41	1	4.6	4.5	4.5
3	Cs	1.7	1.5	0.2	0.18	0.22	40	6	2	4.5	4.5	4.5
4	Da	1.4	1.3	0.17	0.16	0.21	32	21	3	5.2	5.2	5.2
5	He	1.4	1.4	0.14	0.14	0.18	27	13	1	4.4	4.4	4.4
6	Ca	1.3	1.1	0.14	0.15	0.18	30	9	3	5.7	5.3	5.6
7	Es	1.2	1.3	0.17	0.17	0.22	29	9	2	4.9	4.8	4.9
8	Fi	1.4	1.4	0.2	0.2	0.24	37	11	2	4.5	4.5	4.5
9	Hu	1.3	1.2	0.21	0.2	0.24	35	11	2	4.1	4.1	4.1
10	It	1.3	1	0.22	0.2	0.28	27	8	1	4.5	4.5	4.5
11	Nl	1.2	1.3	0.19	0.18	0.23	34	9	1	4.7	4.6	4.7
12	Pl	1.2	1.1	0.17	0.17	0.22	23	8	1	4.7	4.7	4.7
13	Pt	1.2	1.3	0.2	0.17	0.25	30	9	4	4.7	4.7	4.7
14	Sl	1.1	1.1	0.26	0.27	0.32	42	7	1	4.1	4.1	4.1
15	Sv	1.3	1.4	0.14	0.15	0.19	31	9	1	4.7	4.7	4.8
16	Uk	1.4	1.1	0.2	0.28	0.23	39	31	3	5.4	5.5	5.5
17	Zh	1.4	1.4	0.21	0.2	0.25	35	6	3	4.3	4.3	4.3
18	Ja	1.6	1.5	0.18	0.17	0.22	29	4	1	4.1	4.1	4.1
19	Fr	1.3	1.1	0.2	0.16	0.25	23	9	2	4.6	4.7	4.6
20	De	1.4	1.3	0.16	0.17	0.2	30	12	0	4.5	4.5	4.5
21	En	1.3	1.1	0.14	0.13	0.18	-	-	-	4.8		
22	No	1.4	1.3	0.18	0.18	0.21	36	24	12	4.8		
23	Eo	1.4	1.4	0.19	0.16	0.22	40	16	12	4.6		

TABLE III. Network properties of Wikipedia categories in different languages. The total number of category terms is listed in column 3. Number (percentage) of categories corresponding to Wikipedia article is given in column 4 (5). The total number of link and average degree are given in columns 6 and 7. Columns 8 and 9 present the percentage of categories having no corresponding incoming and outgoing links. The percentage of reciprocal links is given in column 10 and the average clustering coefficient in column 11. Column 12 lists the total number of circles longer than 2 (trivial circles of length 2 are listed in column 10). Categories of English Wikipedia are outstanding in respect to circles. However, only 155 unique nodes (out of 51 110) are involved in formation of 178 circles reaching length of 25.

			Categories, associated with articles								
Number	Code	Number of categories	Number	Percentage	Number of links	Average degree	$P(k_{in}=0)$ %	$P(k_{out}=0)$ %	Reciprocal links %	Average clustering coefficient	Number of circles
1	RU	6124	3961	65	7993	1.31	80.4	0	0.1	0.01	0
2	BG	2519	593	24	4149	1.65	50.3	0.9	0	0.02	0
3	CS	1099	465	42	1425	1.3	68.7	0	0.2	0.01	0
4	DA	1815	688	38	2483	1.37	62.6	0.3	0.5	0.01	1
5	HE	2233	947	42	3800	1.7	56.5	0.2	0	0.02	0
6	CA	1264	742	59	1673	1.32	68	0.2	0.4	0.01	2
7	ES	3642	1364	37	5252	1.44	68.1	0.1	0.5	0.02	5
8	FI	1722	698	41	2360	1.37	68.3	0.3	0	0.01	0
9	HU	405	169	42	459	1.13	69.1	5.2	0.4	0.01	0
10	IT	3419	1009	30	5214	1.53	68.9	0.1	0.2	0.01	0
11	NL	4651	2130	46	7938	1.71	62.2	0	0.1	0.01	0
12	PL	3093	1631	53	4808	1.55	67.8	0.1	0	0.01	0
13	PT	2656	1120	42	3649	1.37	66.9	0.4	0.2	0.02	0
14	SL	2963	491	17	3914	1.32	65.3	0.7	0.1	0.02	0
15	SV	4557	1864	41	7148	1.57	68.6	0.2	0.1	0.02	1
16	UK	1013	712	70	1184	1.17	77	0.4	0.1	0.02	0
17	ZH	6154	1945	32	9455	1.54	64	0.3	0.2	0.01	9
18	JA	5866	2391	41	9429	1.61	73.3	0.2	0.2	0.03	7
19	FR	6334	2705	43	10305	1.63	65.9	0.6	0.2	0.02	2
20	DE	12503	5413	43	19020	1.52	67.2	0.1	0.1	0.02	1
21	EN	51110	17068	33	91321	1.79	63.6	1.4	0.1	0.02	178
22	NO	3176	725	23	4618	1.45	63.6	0	0.1	0.01	1
23	EO	710	346	49	937	1.32	67.7	1	0.2	0.04	0

nected. We are thus driven to find a method to compute "hierarchical score," i.e., the score that would provide the best hierarchy. We hereby test five distinct scoring algorithms, three of which are based on either betweenness centrality or on the $k_{in}(i)/k_{out}(i)$ ratio while two others are based on the more complex well-known PageRank [18] and HITS [19] iterative algorithms. In addition to the definition of a score, we need a way to validate our results. Luckily, the Wikipedia contains its own validation.

The network of the Wikipedia articles is complemented with categories that are a special type of entry, many of which (41% in average over languages; Table III, column 5) correspond to articles. These categories contain edges to superior (parent) categories (for instance "*physics*" points to "*science*" and "*academic disciplines*"). These entries are also written and edited by the Wikipedia readers, but are built with the well-defined goal of representing hierarchy. The category network structure approaches an acyclic graph, with almost no circles (Table III, column 12). The lack of circles can be associated with the negligible number of reciprocal edges and the practically null clustering coefficient (Table III, columns 10 and 11). This is actually surprising given the fact that nobody supervises the creation of these categories. Given the complementary categories network, we have validated each of the five tested hierarchy algorithms by comparing the relative position in the hierarchy of every two nodes with a connecting edge in both the articles network and the categories network. We say that the prediction is "correct" if the relative position of the two nodes in the predicted hierarchy corresponds to their related position in the hierarchy dictated by the categories network. In other words, we compare direction of edges connecting pairs of nodes in the obtained and given hierarchies. The category network is obviously far from being full. The validation is thus only valid for the fraction of node couples and connected in both the category and the article networks. Moreover, not all these node couples will have a relative position in the computed hierarchy, since their score may be higher or lower than the

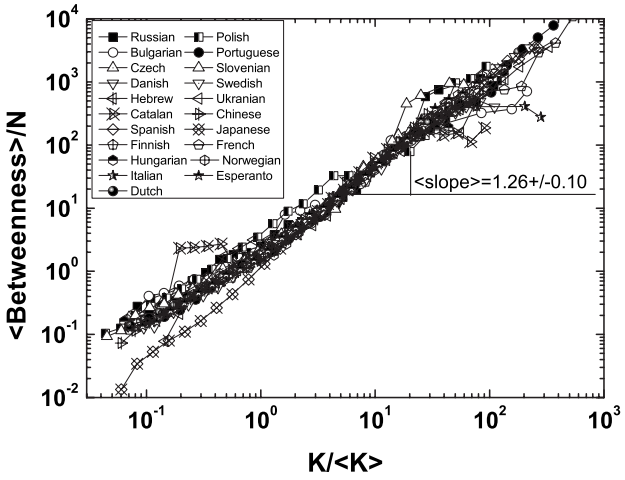

FIG. 5. Average betweenness centrality as a function of node degree collapse to the same function when betweenness is normalized by the number of articles and degrees—by the average degree. The scaled function can be well approximated by power law.

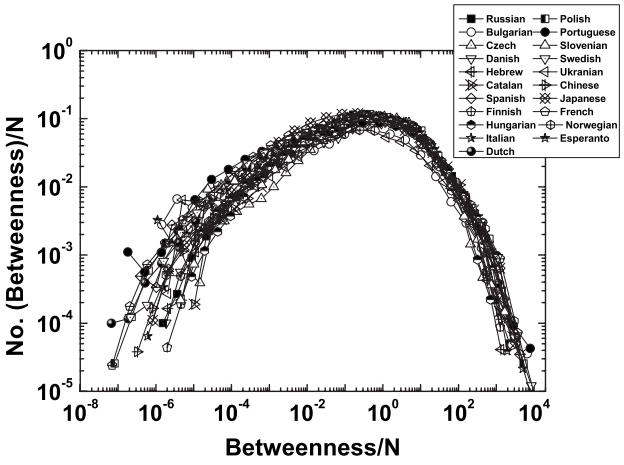

FIG. 6. Scaled betweenness centrality histogram. Betweenness scales well with the network size for all tested languages.

upper or lower thresholds, respectively. Note that the direction of the edge between two nodes in the article network is irrelevant. In other words, if Alabama and the U.S. are connected in both the category network and in the article network, we check that U.S. is positioned above Alabama in the hierarchy.

A. Hierarchical intermediacy

One can develop a hierarchy based on existing network structure. Consider a complete undirected tree with N nodes where each node (unless it is a leaf) has n sons. In a tree there is a single path between every two nodes (since there are no circles). If a node i is above a fraction of p_1 of the nodes, all the paths from a fraction of $(1-p_1)$ to a fraction of p_1 of the nodes and back pass through it (ignoring the node i itself). Contribution of these paths to the betweenness centrality of the node is thus, $C_B(i)=2N(N-1)(1-p_1)p_1$. Furthermore, every node between one of the branches under the node i to another branch must also pass through i, producing a total centrality of $C_B(i)=2N(N-1)(1-p_1)p_1+N(N-1)\times\left(\frac{p_1}{n}\right)^2 n(n-1)=N(N-1)\left(2-p_1-\frac{p_1}{n}\right)p_1$. Given the fact that the fraction of nodes under a level j is n^{-j}, one obtains $C_B(i)=N(N-1)(2-n^{-j}-n^{-j-1})n^{-j}$, which is decreasing exponentially with j, except perhaps for the first level if $n=2$. A similar result can actually be obtained in directed trees, where all edges are directed towards the root node. It would seem from this example that the betweenness centrality is a direct measurement of hierarchy. However, the tree example is artificial, since all nodes (except for the root and the leaves) have equal degrees (both incoming and outgoing). When the degree can vary, the betweenness centrality is proportional to the location, but also to the in degree and out degree (Fig. 5). The first candidate that we have tested for a hierarchy score is thus the centrality $C_B(i)$ normalized to the in degree and out degree. We call this score the hierarchical intermediacy,

$$I(i)=\frac{C_B(i)}{\sqrt{[k_{\text{in}}(i)+1][k_{\text{out}}(i)+1]}}. \tag{1}$$

The +1 term is meant to avoid divisions by zero in the case of no incoming or outgoing edges. The target of this measure is to check whether the node's participation in information flow on the network (betweenness) is higher or lower than what is expected merely from its connectivity. In Wikipedia, the betweenness spans approximately 11 orders and once normalized for the system size, has a very similar distribution among all languages (Fig. 6). We have validated this hierarchal rank measure by comparing the resulting hierarchy in the Wikipedia articles to the explicit hierarchy of the Wikipedia categories. In order to simplify the validation, we have not used a lower threshold. One could get better results by using two cutoffs, but even a single cutoff is enough to produce reasonable results. It turns out that the classification of articles based on the hierarchical rank $I(i)$ is in good correlation, with an average of 82%. Note that the correlation can be as good as 97% (for the Russian Wikipedia) to 62% (for the Norwegian Wikipedia) (Table IV, column 3 and Fig. 9). This fit is good, although not perfect.

B. Local hierarchy

The main limitation of this score is the need to compute the betweenness centrality of the network nodes, which is relatively CPU costly. For very large networks, it is unfeasible to compute the centrality, even using the most efficient algorithms available today [$O(nm)$ [17] for sparse networks, where n is a number of nodes in the network and m is the total number of links]. Moreover, the betweenness centrality is computed for the entire network even if the hierarchy for a small subset of nodes is required (Fig. 7). Another possible flaw is the large contribution of distant nodes. In realistic hierarchical structures, nodes at a minimal distance of 5 or 6 should contribute much less than nearby nodes. If the structure is very far from being a tree, a local definition may be needed. A candidate for a local score is the ratio between in

TABLE IV. Comparison between percentages of successful predictions in different algorithms for tested Wikipedias. English and German Wikipedias were not processed due to their size, which imposed high computational requirements. Attraction-basin algorithm was not applied to Japanese for the same reason. The results were obtained by comparing scores of each pair of nodes for which the hierarchy was defined in the categories' network. Pairs of nodes, for which the ratio between their scores was found to be below some threshold, were considered as indeterminate. Percentage of such pairs is listed in the second column for each algorithm. Thresholds for each algorithm were selected so that the best results would be obtained while keeping percentage of indeterminate node pairs as low as possible (Fig. 9). For simplicity, lower thresholds were not applied.

		Hierarchical intermediacy		Local hierarchy		Attraction-basin hierarchy		PageRank		HITS	
Number	Code	Success	Indeterminate	Success	Indeterminate	Success	Indeterminate	Success	Indeterminate	Success	Indeterminate
1	Ru	97	7	81	32	97	10	95	3	97	25
2	Bg	75	28	86	39	77	31	79	15	73	42
3	Cs	89	30	84	36	85	51	80	21	79	74
4	Da	75	30	81	49	84	42	80	19	79	70
5	He	77	34	90	38	67	40	65	20	42	55
6	Ca	84	30	68	58	80	55	82	19	68	82
7	Es	83	26	80	35	85	39	82	17	85	64
8	Fi	82	33	80	46	84	40	78	17	85	64
9	Hu	94	29	100	65	95	56	97	32	100	94
10	It	84	22	79	37	85	55	86	15	84	70
11	Nl	73	34	78	47	76	49	74	20	80	75
12	Pl	91	21	81	38	94	29	91	9	82	68
13	Pt	88	23	88	29	87	32	84	8	82	51
14	Sl	80	30	91	56	87	55	91	20	89	77
15	Sv	83	23	82	40	91	33	85	14	81	77
16	Uk	95	11	88	22	98	11	98	17	96	22
17	Zh	76	30	79	68	90	50	86	19	89	79
18	Ja	77	32	75	44	-	-	70	25	65	81
19	Fr	79	25	80	47	86	48	84	20	88	70
20	No	62	36	71	37	73	41	72	20	77	73
21	Eo	88	33	82	34	86	49	86	12	82	74
	Average	82	27	82	43	85	41	83	17	81	66

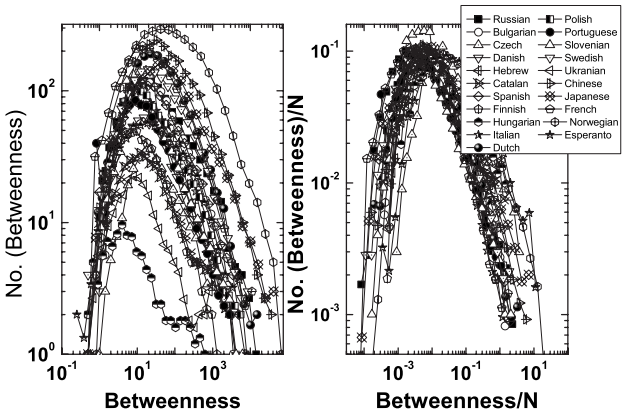

FIG. 7. Betweenness centrality histograms for category networks in different languages (left-hand side). The histogram collapses to the same distribution when scaled with the number of categories (right-hand side).

degree and out degree [Figs. 1(c) and 1(d)]. These ratios exhibit a clear power law, which is practically perfectly conserved among the different languages. Nodes located high at the hierarchy structure tend to be referenced by many "children," while referencing only to a small number of them. If there are only a few references, the excess of incoming vs outgoing edges could just be due to serendipity: having two incoming vs one outgoing is less significant than having 200 incoming vs 100 outgoing. Thus, the relative significance of the score is normalized by the size of the expected fluctuations $(1/\sqrt{k})$. Consequently, we converge to a *local hierarchy* score that is characterized by the following measure:

$$H(i) = \frac{k_{\text{in}}}{k_{\text{in}} + k_{\text{out}}}\sqrt{k_{\text{in}}} + \frac{k_{\text{out}}}{k_{\text{in}} + k_{\text{out}}}\sqrt{k_{\text{out}}}. \tag{2}$$

Note that we have added a symmetrical counterpart to make the score symmetric. While it has the potential to slightly skew the results, the second term has practically no effect in cases where $k_{\text{out}} \ll k_{\text{in}}$. The symmetric score helps to cover other networks that can make use of different direc-

tions to highlight important content. This score performs on the Wikipedia as well as the hierarchical intermediacy (68% to 100%, averaging to 82% in various languages (Table IV, column 5 and Fig. 9).

Though very efficient to compute, the basic caveat of this definition is that it is not sensitive to any underlying network structure. For example, if we would apply this definition to a directed binary tree, this algorithm would fail to find the hierarchy. In the case of a binary tree, there is no difference between any pair of nodes. All nodes, except for the root and the leaves, have one outgoing and two inward edges. We have thus developed a third method incorporating the advantages of the two other methods.

C. Attraction basin hierarchy

A more natural criterion for hierarchy should be based on balancing local and structural considerations. The best such score that we have found is based on the comparison between the weighted fraction of the network that can be reached from each node with the weighted fraction of the network from which the node can be reached (Fig. 8).

More specifically, we define a quantity which summarizes the "flow" of references towards each node i as $\Sigma_m \alpha^{-m} N_{-m}(i)/\langle N_{-m}\rangle$, where $N_{-m}(i)$ is the number of nodes from which the node i can be reached in m directional edges (following shortest paths) and $\langle N_{-m}\rangle$ is the average $N_{-m}(i)$ for all nodes i in the graph. The weight α^{-m} is introduced to balance the influence of the immediate and remote circles. Setting a low value of α enhances the influence of the immediate node surroundings, while high values of α expand the range of influence and make the algorithm more global and structure dependent. The complementary measure, $\Sigma_m \alpha^{-m} N_m(i)/\langle N_m\rangle$, is based on the number of nodes j, located at $m=d(i,j)$ directional edges from the node i. Analogous to the local hierarchy, the resulting attraction-basin hierarchy index is defined for each node i as

$$A(i) = \left(\sum_m \frac{N_{-m}(i)}{\langle N_{-m}\rangle} \alpha^{-m} \right) \Bigg/ \left(\sum_m \frac{N_m(i)}{\langle N_m\rangle} \alpha^{-m} \right). \qquad (3)$$

Simply stated, the algorithm does not only consider immediate nodes' neighbors (like the local hierarchy algorithm), but assigns some (decreasing with distance) weight to distant nodes' neighbors. The advantage of the attraction-basin criterion over the simple local criterion is that it combines both local and structural network properties. Consequently, when applied to a tree, it precisely reproduces its structure.

The hierarchy based on the attraction-basin ranking produced the best fit between the category and article hierarchies (Fig. 9). When applied directly to the category layer with $\alpha=2$, it reproduced the proper structure better than other scores (85%) (Table IV, column 7). Note that as we lower the upper threshold, we reduce the fraction of node couples for which a hierarchical position can be obtained, and increase the success rate for the remaining node couples. The attraction-basin hierarchy produces the highest score and the highest fraction of node couples with a defined hierarchy position (Fig. 9). Note that if one uses a very low cutoff (and consequently a low fraction of classified couples), the betweenness-based hierarchy score described earlier is the one which provides the best score and approaches 100% successful classification.

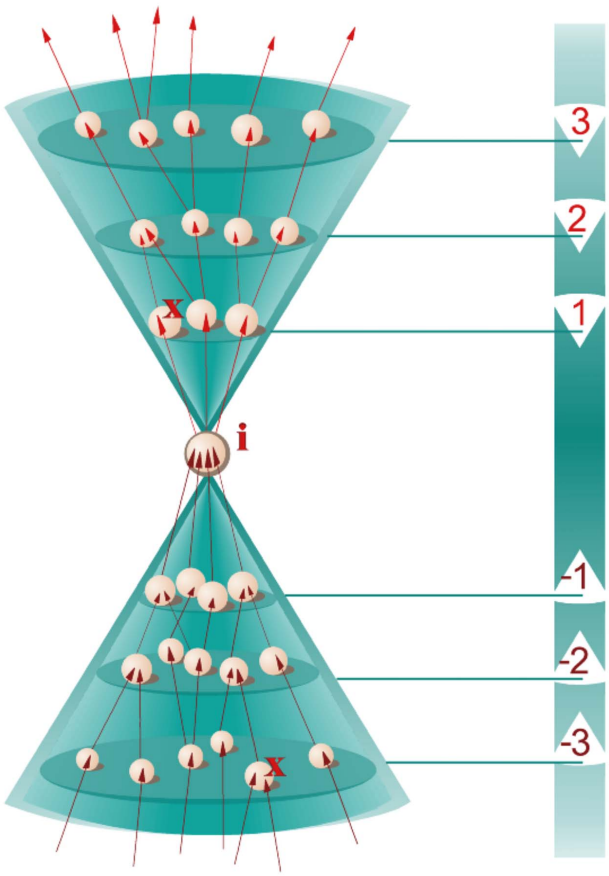

FIG. 8. (Color online) Illustration of the attraction-basin hierarchy algorithm. For each node i, the network is arranged according to each node's distance from it $[d(i,j)]$ such that the node's neighbors are positioned at level 1, nodes at geodesic distance 2 from i at level 2, etc. Simultaneously, nodes j from which i can be reached in m steps $[d(j,i)=m]$ are positioned at level $-m$. Notice that each node could appear 2 times: at one of the levels above and below i. In directed networks, the level numbers may differ (e.g., node marked with x on the graph) since $d(i,j)$ may differ from $d(j,i)$.

We have also tested two of the best-known scores currently used: PageRank [18] and HITS [19]. These two algorithms are widely used, for example, for the classification of Web pages. The disadvantage of these algorithms is that they are iterative and thus have relatively high computational cost. Furthermore, in contrast with the last two scores proposed here, both HITS and PageRank require computing scores for the entire network even if the relative position of a single node couple is desired. The HITS score demonstrates lower success than all other proposed hierarchies, while the PageRank scores similar to the other algorithms proposed here, achieving an 84% success rate at 50% coverage compared with 85% for the attraction-basin hierarchy. Note that all studied hierarchy scores differ in terms of their response

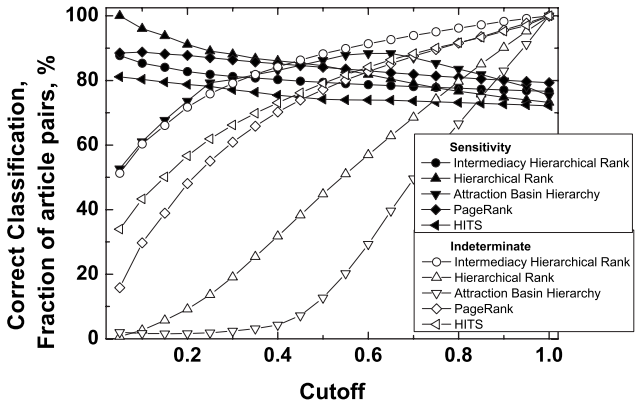

FIG. 9. Sensitivity and fraction of classified links using different algorithms. Black symbols show sensitivity (percentage of true positive classification) of each of the five hierarchy algorithms we have tested as a function of upper cutoff. Cutoff specifies the maximal ratio between the hierarchy scores of the pair of articles which allows determining their respective hierarchy. Hierarchy within pairs of articles having their score ratio above the cutoff cannot be determined and they are removed from the analysis. The fraction of qualified pairs is shown by the hollow symbols. For simplicity, we did not apply lower cutoff, though this degree of freedom can only improve qualification.

to cutoff, influencing the fraction of categorized edges and the overall sensitivity (Fig. 9). Generally, low upper cutoff leads to a stringent definition of the hierarchy, with practically all edges in the proper direction, but with a low number of categorized edges. High cutoff leads to a hierarchy that is often unnatural. A summary of the results obtained for the above-mentioned scores is presented in Fig. 9.

Following the verification of the five proposed hierarchical scores on the number of Wikipedia-induced networks, we test them to other networks containing information that can be represented hierarchically. We have tested three different external networks, one containing a natural hierarchy and two with *a priori* no order. The first network to be tested was a network composed of all DLLs (dynamic link libraries) in the Microsoft Windows XP Pro operating system. Two DLLs are assumed to be linked if direct library calls from one DLL to another exist. Notice that the links are directional and that the obtained network is practically acyclic, since recursive invoking is very uncommon. None of our algorithms require information on the content of the network nodes. Therefore, they can be easily tested with the network of dependencies between software libraries composing the operating system. Using the attraction-basin hierarchy, 6869 links out of 6899 (99.57%) were marked in the proper direction. Next was the local hierarchy producing 98.57% of properly computed links, followed by the PageRank with 98.13%. The centrality-based hierarchy and the one based on HITS had the worst scores (63% and 90%, respectively). Thus, similar to the Wikipedia, the attraction basin and the local hierarchy produced the best scores followed by PageRank.

An interesting case is genetic regulatory networks. Although there is no *a priori* reason for a global hierarchy in these networks, local hierarchical small scale motifs were observed [28]. Each such motif represents a given function of the *E. Coli*, and integration of all such motifs can create a hierarchy. We have applied our attraction-basin hierarchy algorithm to the *E. Coli* regulatory network, and indeed reproduced nearly all the motifs observed by Shen-Orr *et al.* (These figures are large and complex. They can be accessed via Ref. [29]).

Furthermore, all the separated structural motifs were integrated into an acyclic graph where the interaction between the different functions could be observed.

The last tested network, with no *a priori* order, was a neural network. We have applied the first two of the described algorithms to the fully mapped *C. Elegans* neural network, and obtained two different hierarchies. The local hierarchy score reproduces the information flow from sensory to motor neurons [30], while the centrality-based algorithm reproduces the spatial organization from the center outward [31]. These differences highlight the fact that, in the examined neural network, several hierarchies exist simultaneously, each based on different criteria.

Beyond the general statistic properties of the resulting hierarchies, we have inspected a subtree obtained from the hierarchy by selecting one node and picking iteratively only nodes below it. Figure 10 shows a subtree rooted at the term "physics" when the attraction-basin algorithm was used to produce the hierarchy. The first surprising result is that the tree reaches a maximal size and then rapidly decreases. Actually, when the full acyclic graph is taken into account, the number of nodes at each level of the hierarchy decreases as one goes down in hierarchy (in contrast with the naive conception). This decrease results from the fact that all suggested algorithms are based on the information on a given node couple which is gathered from the nodes surrounding the couple. The couples high in the hierarchy are usually thoroughly covered by the Wikipedia users and are consequently surrounded by many nodes (measured by the degree, centrality, or size of their attractors, depending on scores). These nodes have relatively precise information that can vary over a few orders. Examining the difference between these node scores allows us to provide a position in hierarchy. As we go down the hierarchy, the fraction of node couples that can be assigned an order decreases sharply. Thus, although the fraction of nodes at the lower part of the hierarchy should be maximal, we fail to classify most of these nodes. In other words, as we go down the hierarchy, the amount of information on each node couple decreases.

It is important to notice that this is not merely an artifact of our algorithmic approach; rather, it reflects the intrinsic nature of the way the knowledge is arranged. Indeed, although the majority of people would agree upon hierarchical inter-relations between terms appearing high in the hierarchy, our opinions diverge when we approach more specific terms at lower hierarchy levels.

IV. DISCUSSION

We have presented and validated five different algorithms for constructing hierarchy in networks. Two are based on existing Web page ranking methods and three are new algorithms. All proposed algorithms are based on the notion of

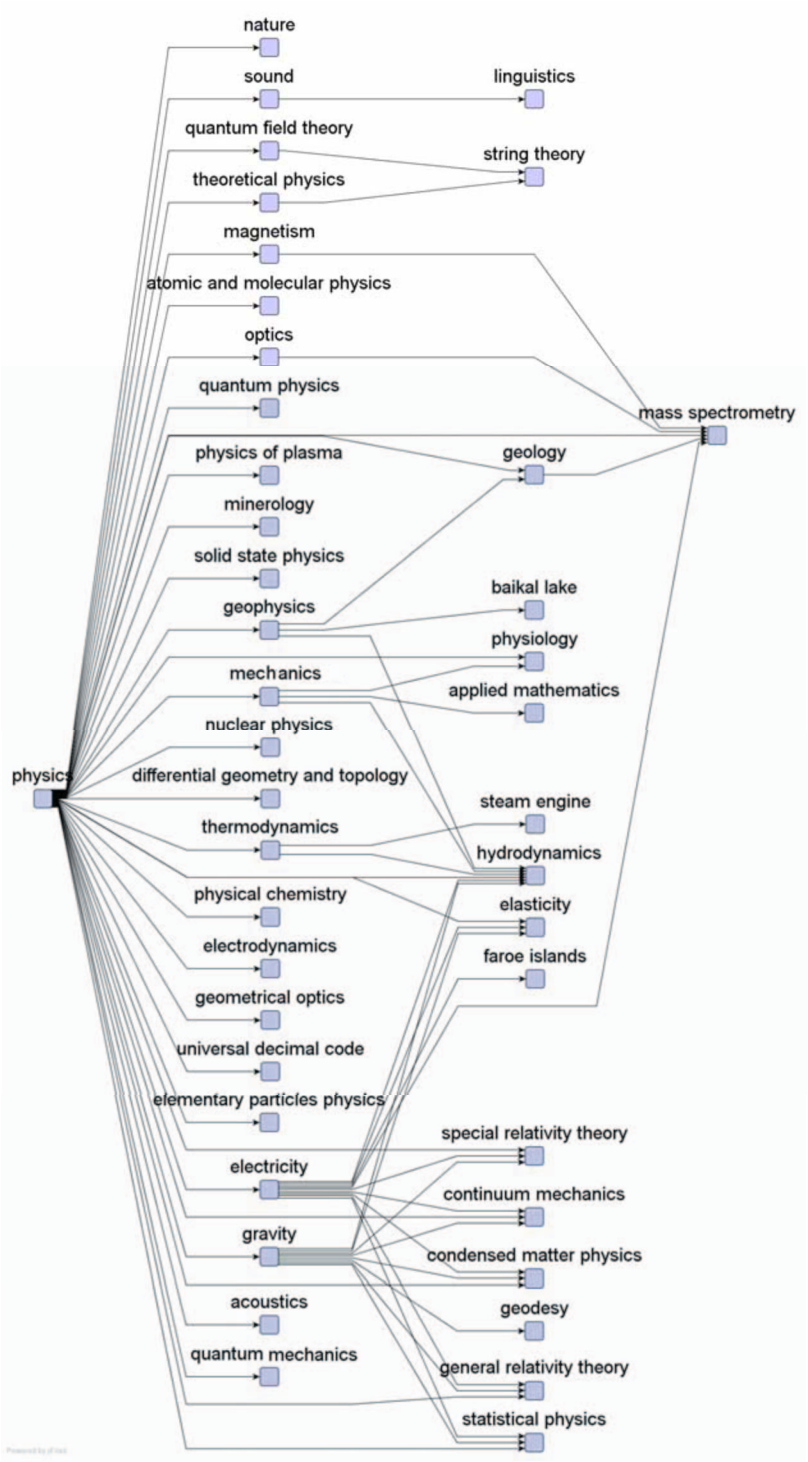

FIG. 10. (Color online) Tree, rooted at the term "physics," automatically generated by application of the attraction-basin hierarchy algorithm to one of the Wikipedias (Russian). The map shows sensible reconstruction of the structure. Those results were obtained with the following set of parameters: $\alpha=2$, upper cutoff -0.8 and lower cutoff -0.1.

information flow and the assumption that it is directed along the underlying conceptual formation. A node is highly positioned in a given hierarchy if it is used to accede to a large number of nodes. The different measures vary by the importance given to distant vs proximal nodes and the weight given to each node. The incoming to outgoing degree ratio is the simplest measure taking into account only first neighbors and assigning to them equal weights. Only first neighbors are used also in the Web ranking methods (based on PageRank and HITS), but different weights are assigned to different nodes. The weighting of neighboring nodes forces them to use iterative methods and to measure the rank of the entire network, even when looking at a small number of nodes. The betweenness centrality-based measure assigns equal uniform weights to all nodes, while the attraction basin assigns nodes at decreasing weights as a function of their distance.

These measures do correspond to the natural way we comprehend information, by relating it to the surrounding pieces. Simply stated, a concept is higher in hierarchy if it squeezes more tightly the information flow in the network.

Interestingly, these measures, related to information flow, are perfectly conserved among Wikipedias in all tested languages, showing that they are indeed an inherent measure of the concept network structure and are not affected by the language dependent representation.

Note that all the proposed algorithms are statistical, and as such provide only a probable hierarchy. This is highlighted by the fact that none of the algorithms can reach a 100% overlap with the category network in the Wikipedia. Moreover, the network itself contains a large number of wrong or doubtful links. Even if the algorithm was perfect, these links would induce a basal error level. We attempt to reduce the error level to the minimum possible, given the random aspects of the network. On the other hand, even human-designed categories are not perfect, and may contain errors as well as controversial relations. Thus, even if the algorithm was perfect, we would expect a basic error level.

ACKNOWLEDGMENTS

The work of two of the authors (L.M. and Y.L.) was covered by the Yeshaya Horowitz. The work of three of the authors (S.S., Y.L., and R.I.) is also supported by the co3 NEST PATHFINDER of the EU 6th framework. Two of the authors (S.S. and L.M.) acknowledge support from DAPHNet FET of FP6.

[1] Aristotle, Metaphysics, Vol. IV, p. 1.
[2] Encyclopédie, Encyclopédie ou dictionnaire raisonné des sciences, des arts et des métiers, Editions Flammarion.
[3] P. Blom, *Enlightening the World: Encyclopedie, The Book That Changed the Course of History* (Palgrave Macmillan, New York, 2005).
[4] B. A. Huberman and L. A. Adamic, Nature (London) **401**, 131 (1999).
[5] R. R. Larson, 1996 Annual ASIS Meeting, Baltimore, 1996.
[6] T. Berners-Lee *et al.*, Commun. ACM **37**, 76 (1994).
[7] Wikipedia http://wikipedia.org/
[8] Pubmed http://www.ncbi.nlm.nih.gov/pubmed
[9] Wikimedia http://www.wikimedia.org/
[10] A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli, Phys. Rev. E **74**, 036116 (2006).
[11] V. Zlatic, M. Bozicevic, H. Stefancic, and M. Domazet, Phys. Rev. E **74**, 016115 (2006).
[12] M. Faloutsos, P. Faloutsos, and C. Faloutsos, Comput. Commun. Rev. **29**, 251 (1999).
[13] Wikipedia categories http://en.wikipedia.org/wiki/Wikipedia:Browse
[14] http://download.wikimedia.org/wikipedia/
[15] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).
[16] L. C. Freeman, Sociometry **40**, 35 (1977).
[17] U. Brandes, J. Math. Sociol. **25**, 163 (2001).
[18] L. Page *et al.*, Stanford Digital Libraries Working Paper, 1998.
[19] J. Kleinberg, *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms* (ACM, New York, 1998), p. 668.
[20] D. R. Hofstadter, *Godel, Escher, Bach: An Eternal Golden Braid* (Basic Books, New York, 1979).
[21] F. Heylighen, *Foundations of Science* (Springer, Amsterdam, 2000), Vol. 4.
[22] A. L. Barabasi and R. Albert, Science **286**, 509 (1999).
[23] P. Erdos and A. Renyi, Institute of Mathematics Hungarian Academy of Sciences, 1959.
[24] L. A. Amaral *et al.*, Proc. Natl. Acad. Sci. U.S.A. **97**, 11149 (2000).
[25] S. Strogatz, Nature (London) **410**, 268 (2001).
[26] H. Jeong *et al.*, Nature (London) **407**, 651 (2000).
[27] Y. Louzoun, L. Muchnik, and S. Solomon, Bioinformatics **22**, 581 (2006).
[28] S. Shen-Orr *et al.*, Nat. Genet. **31**, 64 (2002).
[29] http://peptibase.cs.biu.ac.il/wiki/Data/EcoliGenNetwork/EColiHierarchy.pdf
[30] http://peptibase.cs.biu.ac.il/wiki/Data/CelegansNeuralNetwork/NeuralNetHierarchy1.pdf
[31] http://peptibase.cs.biu.ac.il/wiki/Data/CelegansNeuralNetwork/NeuralNetHierarchy2.pdf