Curvature of Co-Links Uncovers Hidden Thematic Layers in the World Wide Web

Jean-Pierre Eckmann[1,2] **& Elisha Moses**[3]

[1]*Dépt. de Physique Théorique, Université de Genève, CH-1211 Genève 4, Switzerland*
[2]*Section de Mathématiques, Université de Genève, CH-1211 Genève 4, Switzerland*
[3]*Department of Physics of Complex Systems, The Weizmann Institute of Science, Rehovot 76100, Israel*

Beyond the information stored in pages of the World Wide Web, a type of "meta-information" is created when they connect to each other. This new information is a collective effect of independent users writing and linking pages, hidden from the casual individual user. Accessing it and understanding the inter-relation of connectivity and content in the WWW is a challenging problem[1,2,3,4]. We demonstrate here how thematic relationships can be precisely located by looking only at the graph of hyperlinks, gleaning content and context from the Web without having to read what is in the pages. Noting that reciprocal links (co-links) between pages signal a mutual recognition of authors, we focus on triangles containing such links, since triangles indicate a transitive relation. The importance of triangles is quantified by the clustering coefficient[5] which we show is actually a curvature[6,7]. This defines a Web-landscape whose connected regions of high curvature characterize a common topic. We show experimentally that reciprocity and curvature together capture this meta-information, for a wide variety of topics. As an example of future directions we analyze the neural network of *C. elegans*[8,9], using the same methods.

The Web is a graph[10,11] that is continuously being expanded by an enormous number of *independent* agents, who add pages in an uncoordinated way, but obey clear rules on how to address other pages[12]. Their unified action creates a *cooperative meta-information* whose locus is in the connectivity of the network (in contrast to a "hermit" archive of information).

This meta-information has been noted before[2,3], and its utility for data mining explored in the influential works[13,14]. This approach views a link as conferring "authority" and searches the WWW for "Authorities" and "Hubs" that point at them. Based on these ideas the successful search engine Google was designed; it also lead to the notion of "Community"[15,16]. Similar investigations of the information content of linkages are common in other fields, such as studies of social[17] or paper citation networks[18].

The billions of pages and links of the WWW create a maze of confusing and almost unmanageable complexity. To navigate it and to find pages of interest, users therefore rely on the help of search engines[19]. These send out in turn small automated programs (robots) that roam the Web, retrieve links and note keywords as they create a mirror image of the Web, that can be stored within one room of storage data, ready for easy access. The Web remains intelligible to its human users only because it is constantly analyzed and monitored by these automatic agents. Robot activity dominates Web traffic and accounts for up to 70% of the total queries made at certain sites we studied. Because of the breadth of their search, robots obtain a clearer view of the Web than the common user.

We therefore adopt the perspective of robots to get a hold of the meta-information, without having to define meaning or content, looking only at the identification tag of pages (URL addresses) and their links[20]. To avoid the difficulty of having to know the complete Web graph we only use some locally visible pieces which can be found by starting from a given page and progressing along its links. An analogy may be drawn with classical neurobiology[21], where a neuron takes up dye which then traces out its connections.

Having defined our aim of revealing meta-information in the form of hidden thematic structures, we implement it by a novel combination of a number of key ideas: **Clustering, co-links, triangles, and curvature**. The clustering regroups a "home" page with the pages in its sub-directories. Th co-links are then the result of independent agents linking their pages. Triangles capture transitivity, which is measured by the associated notion of curvature. Using it, we show that practically all the connective information depends on co-links. Thus, a new geometric view of the locus of information in networks appears. Note that this information is not "put there" but "it happens" as a collective effect which our method reveals.

The first step, of our procedure is *clustering*[22]. Once a certain number of pages have been found by our robot we regroup them by lumping together a "home" page with the pages in its sub-directories. The advantages are:

- Information within a given home site belongs in one contextual heading.
- The size of the graph is reduced by a factor of $10^2 - 10^3$.
- Connective links that arise from physical proximity are eliminated, leaving the more relevant remote links.

Working henceforth exclusively with the clustered graph (like the one on the right of Fig. 1), we introduce the pivotal notion of **co-link** for a reciprocal connection between two clusters A and B (A points to B *and* B points to A, the red link in Figure 1). They are then deemed to be "congruent" (friends, or members in the same interest group).

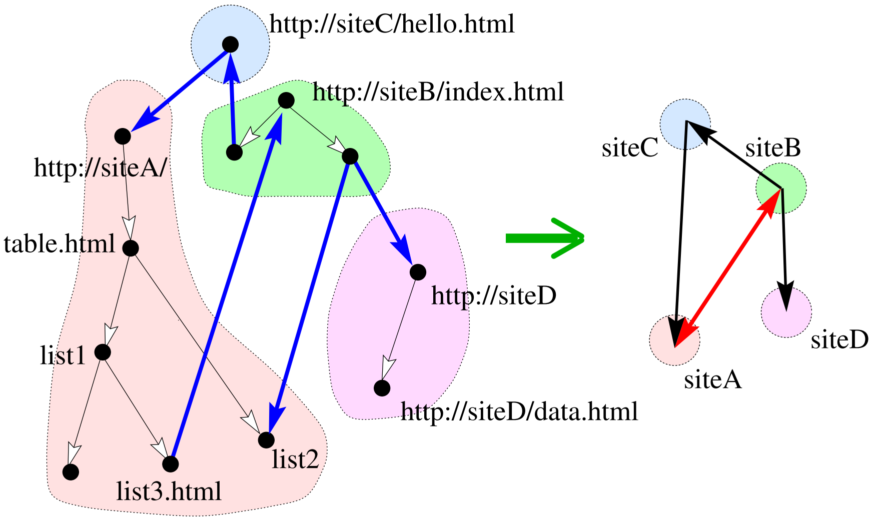

Fig. 1: The reduction of a set of nodes and links by clustering. On the left, local links are black and remote links are blue. On the right, black links are One-way and the red one is a co-link. Clustering clarifies connectivity because it allows the links to go through *different* pages within the clusters.

Co-links are special because, while I can point to your page I cannot (in general) cause you to point to mine. Congruence therefore indicates an awareness by both clusters of the other's existence and content, and a recognition of the other's value to their own interests. Since congruence involves mutual recognition, we expect it to be a transitive property i.e., if sites A and B are congruent, and B and C are congruent, then A and C will be congruent with high probability, forming a *triangle*. Such triangles signal strong *co-operative content*, and will be rare in a random graph[23]. To quantify the clustering of triangles with congruent edges, we define the *local curvature* (introduced previously by Watts and Strogatz as a clustering coefficient[5]) at a node n by $c_n = 2t_n/((v_n - 1)v_n)$. Here t_n is the number of triangles containing n as a corner, v_n is the number of links leaving n (the valence) and $(v_n - 1)v_n/2$ gives the maximal number of possible triangles. We ignore in this the directionality of One-way links. If the graph is a tree, there are no triangles, and $c_n = 0$ everywhere, while for any complete graph (where all nodes are reciprocally connected) $c_n = 1$ everywhere.

To see that c_n is indeed a quantity like curvature, note that it is related to the mean distance (measured by counting links in the shortest path from one node to another) between the nearest neighbors of the node n (the nodes n', n'' at distance 1). Namely, $c_n = 2 - \langle r(n', n'') \rangle$, where r is the distance and the average is taken over all pairs of nearest neighbors, and the average triangle has sides 1, 1, r. On surfaces this relation of triangles to curvature comes naturally as a generalization of the law of cosines relating the angle at the apex and the three sides. It has been generalized[7] by comparing triangles in a continuous metric to their counterparts embedded in standard manifolds. While we are not aware of a corresponding definition for finite graphs with discrete metric, we still view c_n as the curvature of triangles with apex at n (though c_n is normalized to take values between 0 and 1), with c_n being perhaps closer to a dimension measurement[24]. Equipped with a metric and a curvature, the graph of the WWW looks like a hyperboloid when it is "fanning out"[1] or rather more like a sphere when it is "closed" on itself.

Triangles tend to aggregate, creating interest groups that are typically of small diameter (though their size can vary widely). In general, the border of interest groups is expected to be a sharp interface, but sometimes different interest groups will connect one to another, typically via a single link (e.g., the member of the "butterfly collection" interest group who also belongs to the "origami" interest group).

We tested our general ideas in several Web crawls, which are initialized by choosing a first site (the *anchor*) and reading it in. The links in a fetched page are identified, and entered into a database. Our robot then follows all the links in that page and reads in the pages they point to. The database is stored for offline analysis, while intermediate results are used to direct the further action of the robot.

Links are determined to be *local* or *remote* based on a protocol implementing the ideas illustrated in Fig. 1. A local link is basically one inside a given site and any other link is remote. (Details can be obtained from the authors). The algorithm explores a given page (e.g., the anchor) by first following the local links down to depth d (and perhaps u upwards, links like "back to home page"). After exhausting the search down and up the database is reclustered and the robot is sent out along the first remote link, and explores that node, going down and up the newly found local links as before. Once that is done the robot returns to the first node and follows the next remote link, until all remote links at distance $r = 1$ from the anchor have been followed. The robot can then be directed to explore larger distances, if needed.

Large numbers of links in a page (indicating the page is a hub) tend to stall the crawl, and we followed at most 35 links per page, neither were `.cgi, .jpg` explored. As an aside, we note that `.com` sites seem to make little contribution to "cooperative content" and connect badly to others.

Anchor	URL encountered	URL read in	Clusters	Total links	Co-Links
Shakespeare	277,114	69,982	1,560	3,730	321
Needlework	341,398	102,895	1,498	4,440	727
Revisionist	66,771	22,933	947	1,307	67
Piazzolla	47,978	20,404	2,125	2,577	70
Mol. Biol.	318,705	110,286	1,518	6,351	868

Table I: Data collected in 5 experimental Web crawls. Our protocol ensures that nodes within distance $r = 1$ of the anchor are fully explored to a given depth $d = 3$. The crawls started at sites of literary studies of Shakespeare, instructions for needlework, a notorious Swiss revisionist site, tango (specifically the music of Piazzolla), and a site for molecular biology.

The breadth of crawls is shown in Table I. The curvature was calculated for all nodes within a distance $r = 1$ of the anchor, and for the 3 types of graphs defined in Table II. For all experiments, the distribution of curvatures c among the nodes of the Full graphs obeys a power law, $c = (2.0 \pm 0.2)v^{-1}$, where v is the valence of the node. Our main result, shown in Table II, is that the average curvatures of the Congruence and Full graphs are very similar, while that of the One-way graph is lower by a factor ranging between 2.5 and 21. This means that removal of the co-links results in a dramatic decrease of the curvature. **The curvature of the full graph is carried predominantly by triangles containing at least one reciprocal side**. This curvature is not produced by triangles with three co-links alone, because of their small relative number, but rather from triangles that have a combination of co- and One-way links. A graph of these results is shown in Fig 2.

While we consistently avoided any reference to content of sites in the crawl itself, we checked separately in a detailed manual scan of 784 nodes from our crawls, that no less than 75% to 100% of clusters congruent to the anchor indeed relate to the same topic. Going beyond the WWW we examined the efficiency of curvature as a measure of connectivity in the neuronal network of the nematode *C. elegans*, relying again only on linkage data[25] and comparing afterwards to the classic results[8,9]. Using curvature, we immediately identified three of the major neural circuits (amphids, motor neurons of the nerve ring and "other" sensory neurons of the head). The other circuits are either low-curvature (egg-laying and motor neurons of the ventral cord) or absent from the compilation[25] (tail ganglia). The amphids and nerve ring are particularly well separated into connected components of the congruence graph. Furthermore, interneurons (CPU-like neurons) carry more curvature than sensory (input) or motor (output) neurons. Most co-links are within the same functional circuit (72% for links in triangles, 61% for others) while only 42% of the One-way links share this property.

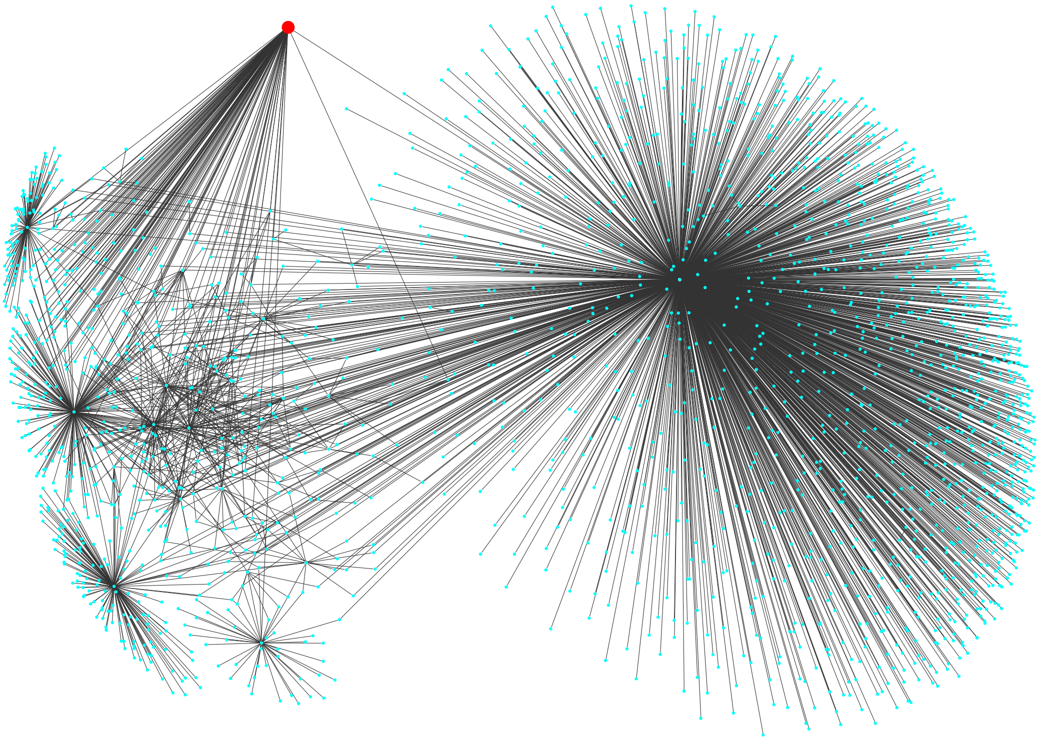

Fig. 2a: 2,125 nodes and 2,755 links of a site specializing in tango (specifically the music of Piazzolla). The full graph.

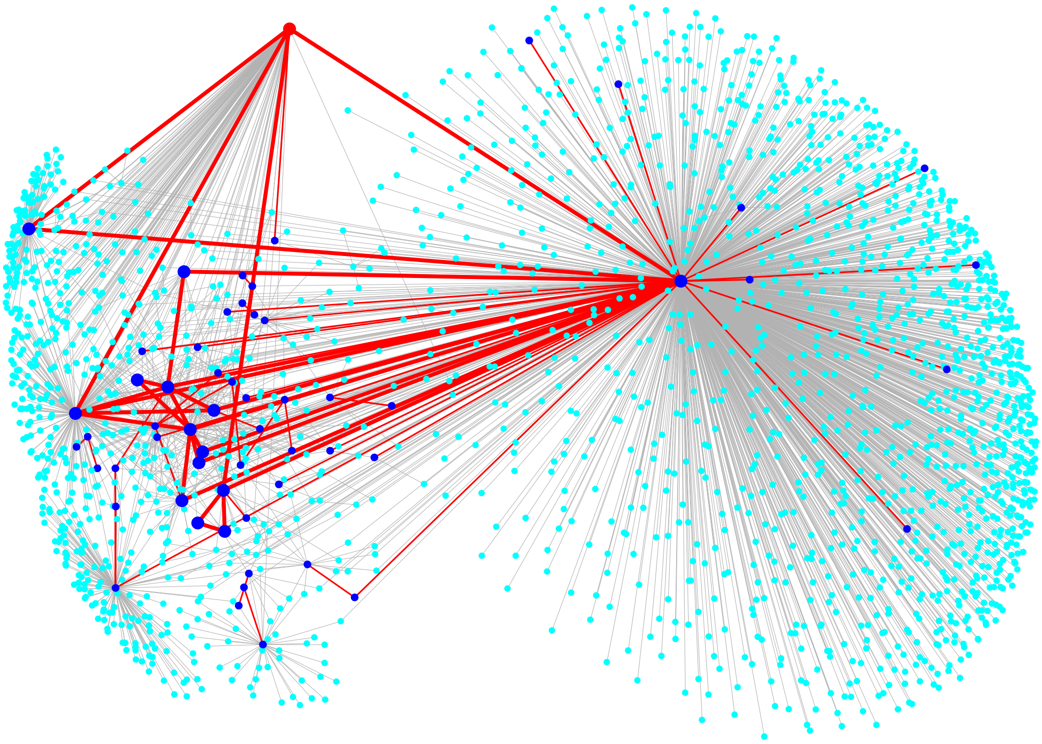

Fig. 2b: The same data with co-links in red, congruence triangles enhanced and one-way links black. The anchor is red. The sparsity and high connectivity of the Congruence graph are striking.

Anchor	Full	Congruence	One-way
Shakespeare	0.46	0.43	0.08
Needlework	0.45	0.54	0.10
Revisionist	0.20	0.20	0.08
Piazzolla	0.29	0.42	0.02
Mol. Biol.	0.50	0.61	0.10

Table II: Average curvatures calculated for three types of subgraphs, using the clusters as nodes and the remote links as edges. The 'Full' graph includes all links that were found in the crawl, the 'Congruence' graph is composed solely of co-links (the red link of Fig. 1), and the 'One-way' subgraph is obtained from the full graph by removing all co-links (leaving the black links).

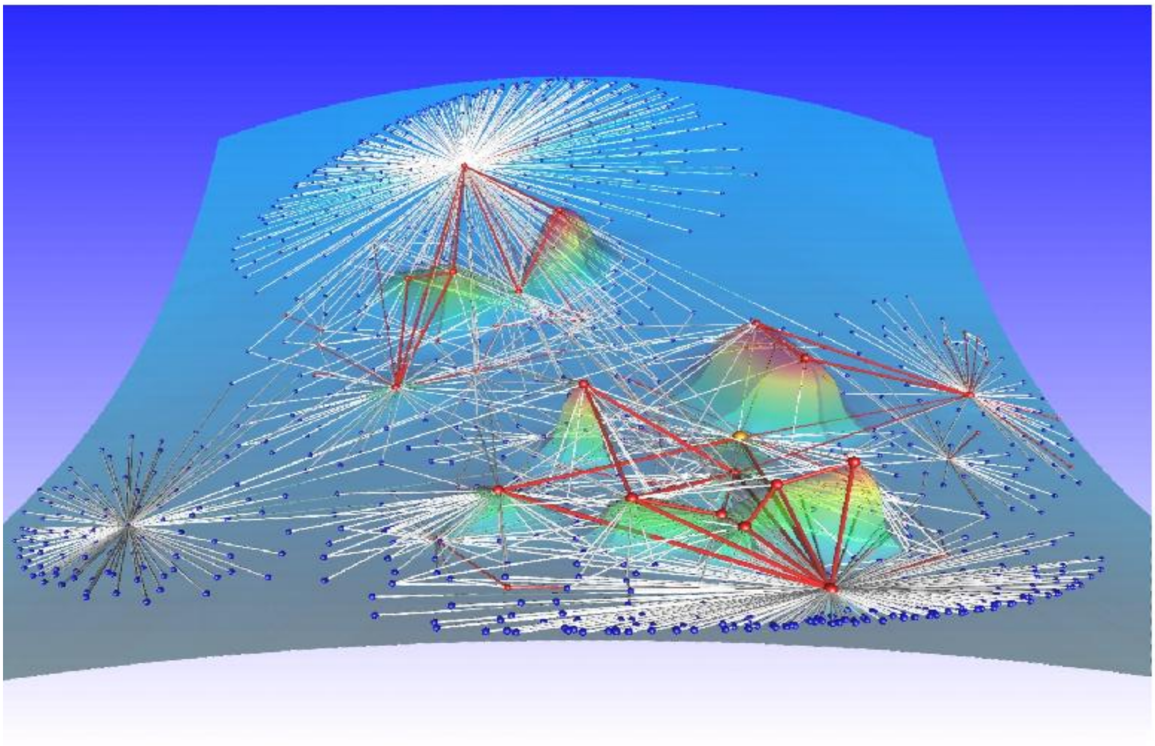

Fig. 2c: A view of the curvature for the 947 nodes and the 1,307 links of a notorious swiss revisionist site. Height above the blue surface is proportional to curvature (of congruence triangles). The One-way graph is clearly less connected, its links are more likely to be dangling ends, and it is generally dominated by the appearance of hubs.

Our discovery opens a number of new possibilities for studying highly assembled complex structures in general, perhaps even the brain. First, the introduction of geometry frees us from the subjectivity of contextual concepts, such as meaning, content, and the like. The locality enables going beyond global notions like fractal dimension of the WWW and the "small-world" effect. Second, the concept of congruence shows the usefulness of combining geometry with selective action in probing sophisticated structures, including *C. elegans*. In particular, we have seen how a very simple dynamical rule – the search for congruence – radically changes our global view of the Web, replacing statistical aspects with unsuspected relations which can be found only by looking at combined, independently built, pathways in the Web. Our approach can certainly be refined and extended in the Web, but should prove useful in many other contexts.

Our study further shows that robots perceive the Web differently than the people who actually write and use the pages. Our robots go back and forth between various sites, gaining a more coherent view of their relations. We have shown that the geometric properties of the space in which they roam and the landscape that they reconstruct reveal new connective meta-information, hidden from the common user.

What can we expect in the future? Other forms of meta-information, not necessarily arising from connectivity, will certainly be found in the WWW and other complex networks. The dynamics of developing interest groups is also an important issue[26,27] that may involve rapid fusion processes[28]. From a mathematical perspective, we have demonstrated the need for concepts of local curvature in graphs. Our definition of curvature seems a useful beginning to elicit Web properties, and is easily generalized to balls of radius $r = 2$ (next-nearest neighbors) or more, as well as to simplices like tetrahedra. But more advanced geometric concepts, measuring non-commutativity in the order of visiting sites, might be needed in the future.

---

1. R. Albert, H. Jeong, A.-L. Barabási, Error and attack tolerance of complex networks. *Nature* **406**, 378–381 (2000).
2. L. Adamic, E. Adar, Friends and neighbors on the Web, preprint (2000).
3. R.A. Botafogo, B. Shneiderman, Identifying aggregates in hypertext structures, *Proc. ACM Hypertext '91* 63–77 (1991).
4. R. Albert, A.-L. Barabási, Statistical Mechanics of Complex Networks, *Rev. Mod. Physics*, submitted. cond-mat/0106096
5. D.J. Watts, S.H. Strogatz, Collective dynamics of 'small-world' networks, *Nature* **393**, 440–442 (1999).
6. M. Gromov, A very accessible exposition of these ideas can be found in the Balzan Lectures of Gromov: 1999. `http://www.balzan.it/english/pb1999/gromov/paper.htm`
7. M.R. Bridson, , A. Haefliger, *Metric Spaces of Non-Positive Curvature*, Grundlehren der mathematischen Wissenschaften, Volume **319** (Berlin 1999, Springer).
8. J.G. White, E. Southgate, J.N. Thomson, , S. Brenner, The structure of the nervous system of the nematode *Cænorhabditis elegans*, Phil. Trans. R. Soc. Lond. B **314**, 1–340 (1986).
9. Wood W.B. ed., *The Nematode Cænorhabditis elegans*, (Cold Spring Harbor Laboratory Press, 1988).
10. J.L. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, A.S. Tomkins, The Web as a graph: measurements, models and methods, *Proc. of the 5th Int. Computing and Combinatorics Conference*, 1999.
11. A. Broder, *et al.* Graph structure in the Web, *Proc. of The Ninth Int. WWW Conference*, 309–320 Amsterdam, The Netherlands, 2000.
12. A.-L. Barabási, R. Albert, Emergence of scaling in random networks. *Science* **286** 509–512 (1999).
13. J.M. Kleinberg, Authoritative sources in a hyperlinked environment, *Jour. ACM* **46** 604–632 (1999).
14. S. Brin, L. Page, The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the 7th International World Wide Web Conference*, Brisbane, Australia, 107–117. 1998, Elsevier).
15. D. Gibson, J.M. Kleinberg, P. Raghavan, Inferring Web communities from link topology. In *Proceedings of the Ninth ACM Conference on Hypertext*, 225–234, (New York 1998, ACM)
16. G.W. Flake, C. Lawrence, C. Lee Giles, Efficient identification of web communities. In *Proceedings of the Sixth International Conference on Knowledge Discovery and Data Mining*, 150–160, (Boston 2000, ACM). Press.
17. S. Wasserman, K. Faust, *Social Network Analysis* (Cambridge 1994, Cambridge U. Press).
18. H. Small, Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, 265–269, (1973).

19. S. Lawrence, C.L. Giles, Accessibility of information on the Web, *Nature* **400**, 107–109 (1999).
20. P. Pirolli, J. Pitkow, R. Rao, Silk from a sow's ear: Extracting usable structures from the Web, *Proc. of ACM SIGCHI Conference on Human Factors in Computing* (1996).
21. R. Cajal, *Degeneration and regeneration of the nervous system* (London, 1928, Oxford University Press)
22. Note this is a different kind of cluster than that described by the "clustering coefficient" of Watts and Strogatz.
23. B. Bollobas, *Random Graphs*, (New York 1985, Academic Press).
24. J.-P. Eckmann, D. Ruelle, Ergodic theory of chaos and strange attractors, *Rev. Mod. Physics* **57**, 617–656 (1985).
25. R. Durbin, `http://elegans.swmed.edu/parts/neurodata.txt`
26. B.A. Huberman, S.M. Maurer, Competitive dynamics of Web sites, preprint (2000).
27. B.A. Huberman, L.A. Adamic, Growth dynamics of the world-wide web, *Nature* **401**, 131 (1999).
28. B. Derrida, S.C. Manrubia, D.H. Zanette, On the genealogy of a population of biparental individuals, *J. Theor. Biol.* **203**, 303–315 (2000).

Acknowledgments We are grateful for numerous discussions with our colleagues in Geneva and Rehovot, and for special help by A. Haefliger and D. Ruelle. This work was partially supported by the Fonds National Suisse and the Minerva Foundation, Munich.