Eur. Phys. J. B **38**, 211–221 (2004)
DOI: 10.1140/epjb/e2004-00114-1

THE EUROPEAN PHYSICAL JOURNAL B

Correlated topologies in citation networks and the Web

F. Menczer[a]

School of Informatics and Departments of Computer Science and Physics, Indiana University, Bloomington, IN 47408, USA

Received 5 November 2003 / Received in final form 26 February 2004
Published online 14 May 2004 –

Abstract. Information networks such as the scientific literature and the Web have been studied extensively by different communities focusing on alternative topological properties induced by citation links, textual content, and semantic relationships. This paper reviews work that brings such different perspectives together in order to build better search tools and to understand how the Web's scale free topology emerges from author behavior. I describe three topologies induced by different classes of similarity measures, and outline empirical data that allows us to quantify and map their correlations. The data is also used to study a power law relationship between the content similarity between two documents and the probability that they are connected by citations or hyperlinks. Such finding has led to a remarkably powerful growth model for information networks, which simultaneously predicts the distribution of degree and the distribution of content similarity across pairs of documents — Web pages connected by links and scientific articles connected by citations.

PACS. 89.20.Hh World Wide Web, Internet – 89.75.-k Complex systems

1 Introduction

Document networks have many different types of information, which define as many topological spaces. If we focus on the connections between nodes we see a directed network with edges represented by citations between documents, or hyperlinks between Web pages. Researchers in the field of *bibliometrics* [1] have studied such citation networks since the 1960's yielding local similarity metrics such as co-citation and bibliographic coupling, and studying global properties such as clustering and degree distributions. Many of these properties have been rediscovered — along with new observations and insight — through the recent resurgence of interest in the area of complex networks, fueled by the popularity of large decentralized networks such as the Web. Now physicists, mathematicians and computer scientists are studying information networks with the tools of statistical physics and graph theory [2–4].

Other networks can be built from information in document collections. The coauthorship relationship can be used to build edges between nodes that represent authors. Coauthorship networks have been found to possess many of the critical properties of complex networks, such as small-world and scale free degree distributions [5]. The dynamic relationship between citations, coauthorship, and other collaboration networks (e.g., funded projects) in document collections is also being studied to understand how the dynamics of these topologies affect one another [6].

While the above approaches focus on edges, the "nodes" in information networks are rich objects. Documents such as articles and Web pages contain text, which lends itself to similarity measurements and consequently to the study of interesting topological characteristics such as density and clustering. At the simplest level, one can obtain a network by creating edges between documents based on word cooccurrence, then find clusters of related words. This text mining approach is being used to discover unknown relationships between genes, diseases and drugs based on the biomedical literature [7,8].

Researchers in the field of *information retrieval* (IR) have been active for several decades in modeling and analyzing more sophisticated *lexical* topologies generated by words. In the *vector space model* [9] a document is seen as a bag of words (the same applies for any piece of text such as a page, paragraph, or query). The relative frequency of words, rather than their position, is used to extract a statistical representation of the document. One can build a vector space where each dimension corresponds to a possible term. In this space a document is a vector, typically a sparse one. Various steps are often taken to improve on the basic model. These include removing very common noise terms in a *stop list* ("the," "at," etc.) [10], conflating terms into sets of semantically related words (e.g. "student" and "study") by *stemming algorithms* [11] and use of *thesauri* [12], and weighting frequencies to discount terms based on their general abundance. In a common

[a] e-mail: fil@indiana.edu

weighting scheme called TFIDF (*term frequency · inverse document frequency*) the coordinate of a document d corresponding to a term t is computed by multiplying the frequency of t in d by a discrimination factor based on the number of documents that contain t [13,14].

Given the sparsity of document vectors, traditional metric distances such as the Euclidean and other L-norms are inadequate at capturing the relationships between documents because they are biased by document length — two short documents tend to appear more similar to each other that two long documents just because of the many zero-weight elements. Two main approaches are taken to cope with this issue. One is to normalize document length; this has led to the use of *similarity* measures that focus only on the non-zero elements. The other approach is to use statistical dimensionality reduction techniques, such as the popular *latent semantic analysis* in which one extracts the terms corresponding to the principal eigenvalues of the term-document frequency matrix [15]. Other techniques, outside of the scope of this paper, include document representations that preserve the relative positions of words to compute proximity, and semantic ontologies of terms such as wordnet [16].

Applications of these lexical topologies are found in document retrieval (e.g., search engines), filtering (e.g., spam detection), and classification (e.g., topic tracking). The aim of the vector space model and all other lexical topology techniques is to support such applications by approximating semantic relationships — "a document is related to another document" or "a page is relevant to a query" — from lexical ones. The ultimate goal is to build systems that can automatically establish semantic relationships from measurable quantities such as word frequencies. In order to test such systems, IR researchers often ask human subjects to assess the relevance of documents with respect to given queries. We can also resort to collections of documents that have been manually classified by human experts. For example articles may be classified into an encyclopedia's predefined topic tree, or Web pages into directories managed by portals companies. The resulting classification ontologies are networks that define semantic topologies.

From an applied perspective, a fundamental goal of information networks research should be to analyze the relationship between semantic topology and other topologies based on observables such as text and links, or in other words, to infer semantic relationships automatically. This goal is becoming both more important and more difficult due to the popularity, omnipresence, size, and dynamic nature of the Web. If we knew how to quickly identify, among 10 billion Web pages, the five most useful pages for a user based on a query, we could build the perfect search engine.

This paper reviews an empirical body of work in which I have quantitatively related the network topologies derived from citations and hyperlinks with a lexical topology derived by text analysis and a semantic topology derived from human classification of documents. In Section 2 the three topologies are defined formally. Section 3 outlines how lexical and semantic similarity decay across Web links. In Section 4 I report on a brute-force approach used to directly measure and map the correlations between similarity measures in the three topologies. Section 5 summarizes the implications of the empirical observations of Section 4 for modeling the evolution of information networks.

The work reviewed here has not appeared in publications typically targeted at the physics community. Since statistical physicists are taking a leading role in the study of complex networks, including information networks, it is hoped that the methodologies and results reviewed here can foster stronger collaborations between this community and others that are actively studying information networks from both theoretical and applied perspectives.

2 Three topologies

Let us define similarity measures corresponding to lexical, link, and semantic topologies. One can of course define any number of such measures. Here we focus for the three topological spaces on metrics selected on the basis of various criteria: (i) they are already established and widely used in some scientific community, (ii) they are easy to measure from publicly available data, and (iii) they have desirable mathematical properties.

We also assume that a similarity measure σ can be defined from a distance measure δ (and vice versa) using the relationship:

$$\sigma = \frac{1}{\delta + 1}. \tag{1}$$

2.1 Lexical similarity

For *lexical* or *content* similarity let us turn to the vector space model. A document, query, or Web page is represented by a vector $\mathbf{d} = (w_{d,1} \cdots w_{d,N_t})$ where N_t is the number of terms in the collection, i.e. the dimensionality of the space. An element $w_{d,t}$ is called *weight* of term t in document d. There are many weighting schemes used in IR. The simplest option is *term frequency* (TF): $w_{d,t} = f(d,t)$, the frequency of t in d. In Section 1 I discussed TFIDF: $w_{d,t} = f(d,t) \cdot i(t)$ where $i(t)$ is the inverse frequency of t in the collection. Several forms have been proposed for the function $i()$, for example

$$i(t) = 1 + \log\left(\frac{N_d}{N_{d,t}}\right) \tag{2}$$

where N_d is the number of documents in the collection and $N_{d,t}$ is the number of documents in the collection that contain term t [13]. The use of TFIDF requires global knowledge of the collection, which obviously is not available in the case of the Web. In the work reviewed in the next sections I have used either TF of TFIDF weighting, depending on the data available. However, in all cases stop words are eliminated [10] and other terms are conflated using a standard stemming algorithm [11].

Once the vector space representation of documents is established, we can define a *content similarity* between two document vectors $\mathbf{d_1}$ and $\mathbf{d_2}$ as:

$$\sigma_c(\mathbf{d_1}, \mathbf{d_2}) = \frac{\|\mathbf{d_1} \cdot \mathbf{d_2}\|}{\|\mathbf{d_1}\| \cdot \|\mathbf{d_2}\|}. \quad (3)$$

This is the *cosine similarity* function, which is traditionally used in IR because it does not suffer from the dimensionality bias that makes L-norms inappropriate, as discussed in Section 1. It is illustrated in Figure 1A.

2.2 Link similarity

The network topology of hyperlinks or citations (links for short) defines a natural distance metric:

$$\delta_l(d_1, d_2) = \min(|p(d_1 \to d_2)|, |p(d_2 \to d_1)|) \quad (4)$$

where $p(u \to v)$ is the shortest path from u to v (links are directed edges) and $|p|$ represents the length of path p. This distance measure will be used in Section 3. However, it has limitations. In some cases there may be no path, for example between two articles in a citation network. Or there may be no directed path, even if a path exists using undirected edges. In other cases a path may exist but shortest paths may not be computable due to incomplete knowledge of network connectivity. This latter problem is typical for the Web. Even a relatively large sample with millions of pages is likely to contain many pairs of pages for which equation (4) would not allow to define δ_l.

A more localized link similarity measure is therefore necessary. Let us define the link neighborhood U_d of a document d as the set of documents that are linked from d or link to d, plus d itself. We can then define a *local link similarity* from a simple Jaccard coefficient:

$$\sigma_l(d_1, d_2) = \frac{|U_{d_1} \cap U_{d_2}|}{|U_{d_1} \cup U_{d_2}|}. \quad (5)$$

Local link similarity measures the degree of clustering between the two pages. To see why, note that if a page has a high clustering coefficient, then it must have a high link similarity to its neighbors. The measure is illustrated in Figure 1B. A high value of σ_l indicates that the two pages belong to a tightly clustered set of pages. Related measures are often used in link analysis to identify a community around a topic. If $\sigma_l(d_1, d_2) > 0$ there exists an undirected path between d_1 and d_2 of length $\ell \leq 2$ links. The higher σ_l, the greater the probability that there is a directed path between the two pages, which could be navigated by a user or crawler. Note that σ_l is also akin to the well known *co-citation* and *bibliographic coupling* measures used in the bibliometrics community.

2.3 Semantic similarity

The traditional IR approach to estimating the semantic relationship between two objects (e.g., a query and a document) is to conduct a user study, asking subjects to estimate the degree of relatedness between the two objects.

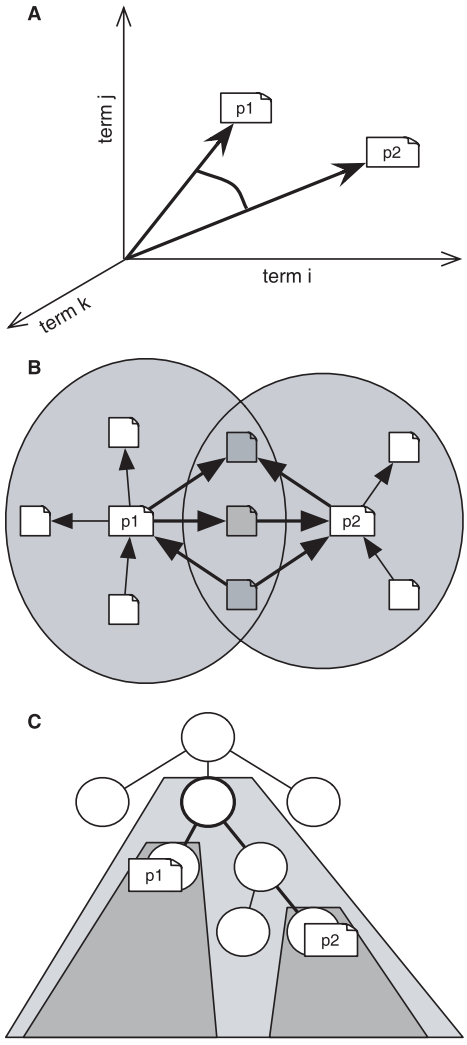

Fig. 1. Illustrations of similarity measures in three topologies. (**A**) Content similarity: the terms shared by the two documents are measured by the cosine of the angle between the two corresponding word vectors. (**B**) Link similarity: the clustering of the two documents is measured by the number of their shared neighbors (dark gray pages), relative to the size of the union of their neighbors (light gray sets). (**C**) Semantic similarity: the meaning shared by the two documents or pages is measured by the entropy of their lowest common ancestor topic (light gray subtree), and the meaning differentiating the two is measured by the entropy of their respective topics (dark gray subtrees).

For example, subjects might be asked to rank a set of documents according to their relevance to a given target query. While users may have their own bias, this is considered the golden standard for IR system evaluation. However, user assessments are very expensive and time consuming for large collections. The approach is infeasible when one needs to consider all pairs of documents in a large set, as is required to measure the correlations between semantic similarity and other similarity measures.

Fortunately, we can rely on large sets of pre-classified documents without renouncing the golden standard of human assessments. Digital libraries are often marked up with descriptors that categorize articles into some ontology. Examples of these include the ACM Computing Classification System, the AIP Physics and Astronomy Classification Scheme, and the NLM Medical Subject Headings and Classification. For Web pages, large directories have been built manually. The simplest version of this idea is a hierarchical taxonomy with pages classified at nodes, which correspond to categories or topics. The best known examples of Web directories are Yahoo[1] and the Open Directory Project[2] (ODP). The latter is maintained by a large number of volunteer editors, makes its data publicly and freely available, and does not have a strong commercial bias — there is no mechanism to pay in order to be listed. These directories are large, with hundreds of thousands of topics and millions of pages. Their ontologies also have more complex structures than a simple hierarchical taxonomy. There are symbolic links between topic nodes in different branches as well as links describing non-hierarchical relationships. These result in complex networks that, unlike trees, have weighted edges and cycles.

In the simple case of a tree ontology, we can define a *semantic similarity* between two documents using the entropy of the documents' respective topics:

$$\sigma_s(d_1, d_2) = \frac{2 \log \Pr[t_0(d_1, d_2)]}{\log \Pr[t(d_1)] + \log \Pr[t(d_2)]} \qquad (6)$$

where $t(d)$ is the topic node containing d in the ontology, t_0 is the lowest common ancestor topic for d_1 and d_2 in the tree, and $\Pr[t]$ represents the prior probability that any document is classified under topic t. This measure is illustrated in Figure 1C. In practice $\Pr[t]$ can be computed offline for every topic t in the tree by counting the fraction of documents stored in the subtree rooted at node t, out of all the pages in the tree. The path from the root to t_0 is a measure of the meaning shared between the two documents, and therefore of what relates them. Conversely the paths between t_0 and the two document topics is a measure of what distinguishes the meanings of the two documents. This semantic similarity measure is a straightforward extension of the information-theoretic similarity measure [17], designed to compensate for the fact that the tree can be unbalanced in terms of both its topology and the relative entropy of its nodes. For a perfectly balanced tree in which all documents are evenly stored at the leaves, σ_s is equivalent to the familiar tree distance measure (normalized length of shortest tree path).

In Section 4 we use the semantic similarity definition of equation (6) for Web pages based on ODP data. Sampling pages from the ODP guarantees that semantic information for each page is available from human editors. However, as discussed above, the ODP ontology is not a simple tree. For example, the "Business" category is subdivided by types of organizations (cooperatives, small businesses, major companies, etc.) as well as by areas (automotive, health care, telecom, etc.). Furthermore, the ODP has various types of cross-reference links between categories, so that a node may have multiple parent nodes and be reachable from the root following multiple paths. How to extend the definition of equation (6) to this graph is the object of ongoing study. In the work reviewed here, the ODP ontology is reduced to a tree by disregarding cross-reference links and other links that disrupt the simple hierarchical topology. This introduces a form of noise into this measure — two Web pages may be more strongly related than the measure indicates.

[1] `http://www.yahoo.com`

[2] `http://dmoz.org`

3 Clustering

In this section I review how lexical and semantic relationships decay across link distance, i.e., how lexical and semantic similarity are autocorrelated in link space [18].

Link distance is defined by equation (4), and shortest paths are discovered by an exhaustive breadth-first crawl. The large fan-out of Web pages imposes a practical limit to the maximum link distance that we can measure. The collection used for these experiments was obtained by starting a a breadth-first crawl from each of 100 topic pages in the Yahoo directory. Yahoo pages were used only as starting points — the crawl was entirely outside of Yahoo.

Lexical similarity is measured by cosine similarity (Eq. (3)) using TFIDF weighting with inverse document frequency (Eq. (2)) computed from the collection of Web pages crawled. Cosine similarity was computed between each crawled page and the name of the topic where the crawl originated.

The choice of starting points for the crawls in a Web directory was driven primarily by the need to measure semantic similarity between crawled pages and starting pages. Even though crawled pages are not manually classified (making it impossible to use Eq. (6)), we can deem a crawled page semantically related to the starting topic if it links to one of the starting pages (which are assessed as highly relevant to the topic by the Yahoo editors). This idea is formalized below.

To obtain meaningful and comparable statistics at $\delta_l = 1$, only pages with at least 5 external links were used, and only the first 10 links for pages with over 10 links. Topics were selected in breadth-first order and therefore covered the full spectrum of Yahoo top level categories. Each crawl reached a depth of $\delta_l = 3$ links from the start page and was stopped if 10,000 pages had been retrieved at the maximum depth. A timeout of 60 seconds was applied for each page. The resulting collection comprised 376, 483 pages. The text of each fetched page was parsed to extract links and stemmed terms.

3.1 Lexical similarity versus link distance

The measurements were aggregated across all pages within a maximum distance $d \in (1, 2, 3)$ from a seed topic, for

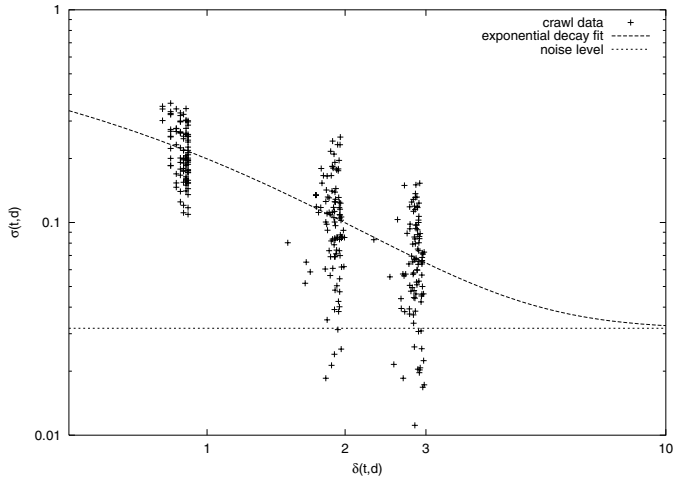

Fig. 2. Scatter plot of $\sigma(t,d)$ versus $\delta(t,d)$ for topics $t = 0,\ldots,99$ and depths $d = 1,2,3$. An exponential decay fit of the data and the similarity noise level are also shown. Data from [18].

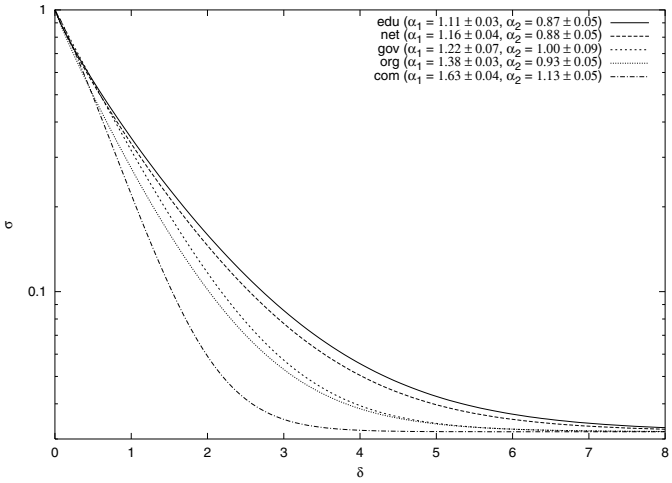

Fig. 3. Exponential decay of $\sigma(q,d)$ versus $\delta(q,d)$ for each of the major US top level domains. The model parameters, obtained via a nonlinear least-squares fit of each domain data, are shown with asymptotic standard errors. For α_1, the differences between `com` and every other domain are statistically significant at the 95% confidence level. Extrapolated from data in [18].

each of the 100 topics t:

$$\delta(t,d) \equiv \langle \delta_l(t,p) \rangle_{P_d^t} = \frac{1}{|P_d^t|} \sum_{i=1}^{d} i \cdot (|P_i^t| - |P_{i-1}^t|) \quad (7)$$

$$\sigma(t,d) \equiv \langle \sigma_c(t,p) \rangle_{P_d^t} = \frac{1}{|P_d^t|} \sum_{p \in P_d^t} \sigma_c(t,p). \quad (8)$$

where $P_d^t = \{p : \delta_l(t,p) \leq d\}$.

The 300 measures of $\delta(t,d)$ and $\sigma(t,d)$ from equations (7) and (8), corresponding to 100 queries $\times$ 3 depths, are shown in the scatter plot of Figure 2. Note that the points are clustered around $\delta_l = 1,2,3$ because the number of pages at distance $\delta_l = d$ typically dominates P_d^t ($|P_d^t| \gg |P_{d-1}^t|$). The two metrics are well anticorrelated (correlation coefficient $\rho = -0.76$). The two metrics are also predictive of each other with high statistical significance ($p < 0.0001$). Such a strong correlation between link and lexical similarity confirms our intuition that authors tend to link pages with similar content.

To analyze the decrease in the reliability of lexical content inferences with distance from the topic page in link space one can perform a nonlinear least-squares fit of these data to a family of exponential decay models:

$$\sigma(\delta) \sim \sigma_\infty + (1 - \sigma_\infty) e^{-\alpha_1 \delta^{\alpha_2}} \quad (9)$$

using the 300 points as independent samples. Here σ_∞ is the noise level in similarity, computed by comparing each topic page to external pages linked from different Yahoo categories:

$$\sigma_\infty \equiv \left\langle \frac{1}{|P_1^{t'}|} \sum_{p \in P_1^{t'}} \sigma(t,p) \right\rangle_{\{t,t' : t \neq t'\}} \approx 0.0318 \pm 0.0006. \quad (10)$$

Note that while starting from Yahoo pages may bias $\sigma(\delta < 1)$ upward, the decay fit is most affected by the constraint $\sigma(\delta = 0) = 1$ (by definition of similarity) and by the longer-range measures $\sigma(\delta > 1)$. The regression yields parametric estimates $\alpha_1 \approx 1.8$ and $\alpha_2 \approx 0.6$. The resulting fit is also shown in Figure 2, along with the noise level σ_∞. The similarity decay fit curve provides us with a rough estimate of how far in link space one can make inferences about lexical content.

The crawled pages were divided up into connected sets within top level Internet domains. The resulting sets are equivalent to those obtained by breadth-first crawlers that only follow links to servers within each domain. The relationship between $\delta(t,d)$ and $\sigma(t,d)$ for these domain-based crawls is plotted in Figure 3. The plot illustrates the heterogeneity in the reliability of lexical inferences based on link cues across domains. The parameters obtained from fitting each domain data to the exponential decay model of equation (9) estimate how reliably links point to lexically related pages in each domain. The parametric estimates are also shown in Figure 3 suggesting that, for example, academic Web pages are better connected to each other than commercial pages in that they do a better job at pointing to other similar pages. Such a finding is not surprising considering the different goals of the two communities. This result can be useful in the design of topic-driven crawling algorithms that prioritize links based on the textual context in which they appear; one could weight a link's context based on its site domain.

3.2 Semantic similarity versus link distance

To see how far semantic signals are carried across Web links, consider the conditional probability that a page p is relevant with respect to some topic t, given that page r is

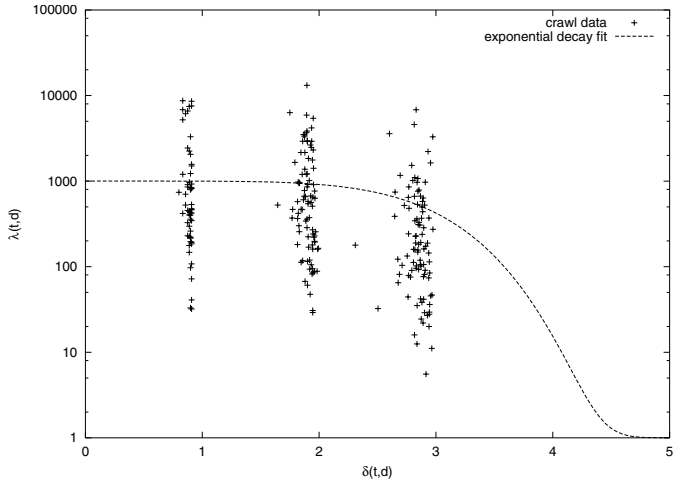

Fig. 4. Scatter plot of $\lambda(t,d)$ versus $\delta(t,d)$ for topics $t = 0,\ldots,99$ and depths $d = 1,2,3$. An exponential decay fit of the data is also shown. Data from [18].

relevant and that p is within d links from r:

$$R_t(d) \equiv \Pr[rel_t(p) \mid rel_t(r) \wedge \delta_l(r,p) \leq d] \qquad (11)$$

where

$$rel_t(p) = \begin{cases} 1 \text{ if } p \text{ is relevant with respect to } t \\ 0 \text{ otherwise.} \end{cases} \qquad (12)$$

$R_t(d)$ is the posterior relevance probability given the evidence of a relevant page nearby. Contrast $R_t(d)$ with the prior probability $G_t \equiv \Pr[rel_t(p)]$, also known as the *generality* of the topic, by defining a *semantic likelihood factor*:

$$\lambda(t,d) \equiv \frac{R_t(d)}{G_t}. \qquad (13)$$

If $\lambda(t,d) > 1$, then a page has a higher than random probability of being about t if it is within d links from other pages on that topic.

To estimate $R_t(d)$ one can use the relevant sets compiled by the Yahoo editors for each of the 100 topics:

$$R_t(d) \simeq \frac{|P_d^t \cap Q_t|}{|P_d^t|} \qquad (14)$$

where Q_t is the relevant set for t. In other words, we count the fraction of links out of a set that point back to pages in the relevant set. For G_t one can use:

$$G_t \simeq \frac{|Q'_t|}{|\bigcup_{t' \in Y} Q'_{t'}|} \qquad (15)$$

where all of the relevant links for each topic t are included in Q'_t, even for topics where only the first 10 links were used in the crawl ($Q'_t \supseteq Q_t$), and the set Y in the denominator includes all Yahoo leaf categories. Finally the measures from equations (14) and (15) were plugged into definition (13) to obtain the $\lambda(t,d)$ estimates for $1 \leq d \leq 3$.

The 300 measures of $\lambda(t,d)$ thus obtained are plotted versus $\delta(t,d)$ from equation (7) in the scatter plot of Figure 4. Closeness to a relevant page in link space is highly predictive of relevance, increasing the relevance probability by a likelihood factor $\lambda(t,d) \gg 1$ over the range of observed distances and queries.

I also performed a nonlinear least-squares fit of this data to a family of exponential decay functions using the 300 points as independent samples:

$$\lambda(\delta) \sim 1 + \alpha_3 e^{-\alpha_4 \delta^{\alpha_5}}. \qquad (16)$$

Note that this three-parameter model is more complex than the one in equation (9) because $\lambda(\delta = 0)$ must also be estimated from the data ($\lambda(t,0) = 1/G_t$). Further, the correlation between link distance and the semantic likelihood factor ($\rho = -0.1, p = 0.09$) is smaller than between link distance and lexical similarity. The regression yields parametric estimates $\alpha_3 \approx 1000$, $\alpha_4 \approx 0.002$ and $\alpha_5 \approx 5.5$. The resulting fit is also shown in Figure 4. Remarkably, fitting the data to the exponential decay model provides us with quite a narrow projection of how far in link space we can make inferences about the semantics (relevance) of pages, i.e., up to a critical distance between 4 and 5 links.

4 Similarity correlations and maps

If we could design maps that, given coordinates based on text and link analysis, told us the position of a document or Web page in semantic space, then we could mine for pages about a certain topic with great accuracy, estimating the meaning of a page from its observable text and link cues — a golden goal for Web mining. This section describes a brute-force approach to map the correlations and functional relationships between the three topologies discussed in Section 2 [19].

As a first step toward charting the semantics of the Web, let us quantitatively analyze the relationship between content, link, and semantic similarity functions across pairs of Web pages. First we want to study whether these different similarity measures are correlated, and secondly we want to ask, given two pages with some lexical and link similarity, what is the likelihood that they are about the same topic.

4.1 Correlations of similarity measures

A set of pages representative of the Web at large was sampled from the ODP, so that semantic information compiled by human editors is available for each page sampled. After filtering out certain parts of the directory tree for language and classification consistency, 10 000 URLs were sampled uniformly from each of 15 top level branches, resulting in a final set of 109,648 URLs corresponding to valid HTML pages in 47 174 topics. The pages were crawled, preprocessed and stored locally for analysis. Then, for each pair of pages I measured their content, link, and semantic similarity as defined in Section 2. Cosine similarity (Eq. (3)) was measured using simple TF weighting. All three similarity measures have values defined in the unit interval. This was divided into 100 bins, resulting in a cube with

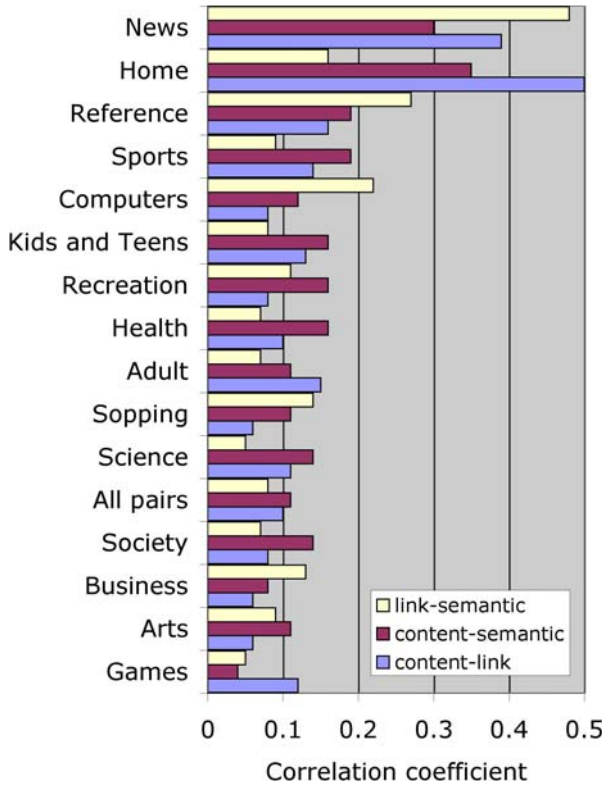

Fig. 5. Correlation coefficients between similarity measures across pairs of pages sampled from the Open Directory. Summary statistics are shown for all pairs and for 15 top level branches of the directory tree.

10^6 bins. Similarity triplets were computed for almost 4 billion pairs of pages from the ODP sample. The data thus collected allows for a number of interesting analyses. Figure 5 shows that there are small positive correlations between all pairs of similarity metrics. Given the very large numbers of pairs, these represent weak but very significant correlations. These numbers quantitatively validate text and link analysis techniques for relevance estimation.

A few exceptionally strong correlations are found, for example in the "Home" and "News" categories. The majority of "Home" sites are about recipes, which often link to related recipes. For "News," it is comforting that journalists seem to use words and links carefully, in a way that helps discern their meaning. These results can be of importance to designers of topical portals and search engines: they indicate which types of analysis are most effective and which topics best lend themselves to specialistic search applications.

4.2 Semantic maps

To visualize how accurately semantic similarity can be approximated from content and link cues, we need to map the σ_s landscape as a function of σ_c and σ_l. There are two different types of information about σ_s that can be mapped for any given (σ_c, σ_l) coordinates: averaging highlights the expected values of σ_s and is akin to the *precision* measure used in IR; summing captures the relative mass of semantically similar pairs and is akin to the *recall* measure in IR. Let us therefore define *localized* precision and recall for this purpose as follows:

$$P(s_c, s_l) = \frac{\sum_{p,q} \delta_c(p,q,s_c)\delta_l(p,q,s_l)\sigma_s(p,q)}{\sum_{p,q} \delta_c(p,q,s_c)\delta_l(p,q,s_l)} \quad (17)$$

$$R(s_c, s_l) = \frac{\sum_{p,q} \delta_c(p,q,s_c)\delta_l(p,q,s_l)\sigma_s(p,q)}{\max_{s'_c,s'_l} \sum_{p,q} \delta_c(p,q,s'_c)\delta_l(p,q,s'_l)\sigma_s(p,q)} \quad (18)$$

where p and q are dummy page indices, (s_c, s_l) is a coordinate value pair for (σ_c, σ_l), and

$$\delta_x(p,q,s) = \begin{cases} 1 \text{ if } \sigma_x(p,q) = s \\ 0 \text{ otherwise.} \end{cases} \quad (19)$$

Note that recall was renormalized by a constant factor for improved visualization.

Figure 6 maps recall and precision over content and link similarity coordinates across all pairs, and for pairs within a few of the top level ODP branches. These semantic maps provide for a detailed signature of the relationship between text, links, and meaning. To properly interpret the recall maps it must be noted that most pairs have small values for all similarity measures (the individual similarity distributions are roughly exponential, each peaked at zero). This makes sense since one would not expect two random pages to be lexically similar, closely clustered, or semantically related. The very small number of pairs with high similarity values explains the weak similarity correlations. Since the majority of pairs occur near the origin, the same holds for most of the semantically related pairs, thus recall is highest near the origin. However all this relevant mass is diluted in a sea of unrelated pairs so that precision near the origin is negligible. This creates an obvious challenge for search engines: achieving high recall costs dearly in terms of precision, leading to user frustration. While emphasis on precision is customary and reasonable for a search engine, the maps reveal how costly this choice is in terms of recall.

The maps also demonstrate that there is significant heterogeneity in semantic landscapes across broad topics. While most of the semantically related pairs occurs near the origin, there are noticeable local optima and ridges in recall that extend away from the origin for several topics. However the recall topology is different for each topic. The topics with higher content-link correlation are those for which more pairs extend away from the origin, and therefore correspond to positive recall values toward high content and link similarity. For topics such as "Home" and "News" it is clear that semantic similarity is correlated with both content and link similarity, making text and links informative cues about page meaning. The "Adult"

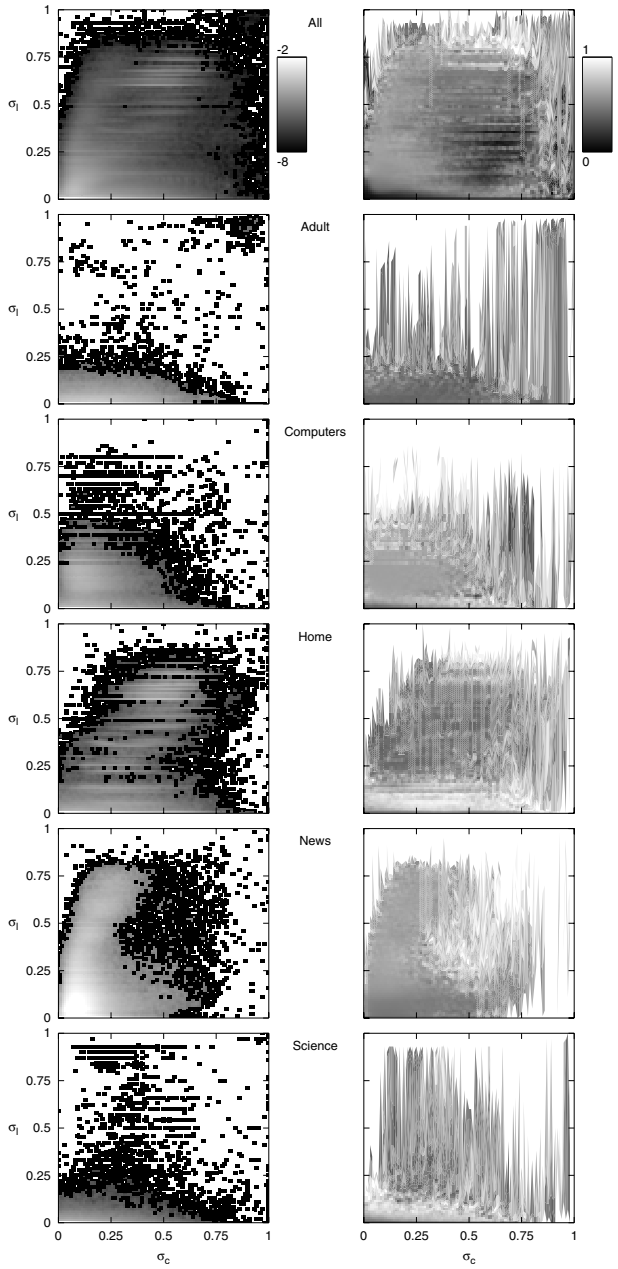

Fig. 6. Semantic maps of recall (left) and precision (right) for pairs of Web pages in the whole ODP sample and within five sample topics. Shades of gray encode the values of recall and precision for each content/link similarity coordinates. Recall is visualized on a logarithmic scale between 10^{-8} and 10^{-2}, precision on a linear scale between 0 and 1. White represents missing data (no pairs).

topic is an exception. There is a large clique of adult sites whose content and links are designed to boost their ranks in search engines such as Google [20]. These engines rank pages primarily by the link-based PageRank metric after selecting pages that contain the query terms. Thus the lonely peak in the top right corner of the "Adult" recall map represents a single business effort rather than an emergent property of independent sites.

On the precision maps one can generally distinguish regions of high precision (shown in light gray) with various sizes, shapes, and locations. The general map shows that a universal search engine should concentrate on the highest link similarity among pages with medium-high content similarity. Surprisingly, for very high content similarity there is significant noise making it difficult to identify relevant pages in this region via link analysis. This sheds light on the low precision of the first generation of search engines, based primarily on lexical similarity metrics, and on the success of the newer generation of engines that exploit link analysis.

Topical precision maps differ significantly from each other and from the general precision map. Most branches have visible regions of high precision. For example several topics such as "Science" have a hot region spanning a wide range of content similarity but a relatively narrow range of low link similarity. The "Home" topic has a second hot region for high link similarity, corresponding to the hot region seen in the general map. A couple of topics ("Computers" and "News") have large, well localized high precision regions. These observations highlight how diverse are the semantic inferences that can be drawn from text and link cues depending on the topical context of a search. These maps also suggest that identifying semantically related pages with high precision is a hard search problem due to many local optima. The optimal strategy for one topic may not be applicable to different domains or to the general case. Simple combinations of lexical and link analysis result in both false positives and false negatives because many high precision regions are isolated and irregular [19].

An important lesson from these maps is that no single approach will work best in the topical context of every user's information need. Search engine companies tend to maintain a universal user base rather than focus on specialized niche domains where the advertising revenues would be smaller. Yet the semantic maps suggest that efforts would be more fruitful if directed at supporting distributed, topic specific search services.

5 Growth models

Another way to visualize the connections between content and link information is to project the similarity data cube onto one or two of its topological dimensions. In this section I review the functional relationship between the probability that two documents are linked, and their lexical distance [21]. This relationship has motivated a growth model for document networks that generates accurate predictions for both link and content distributions in both scientific articles and Web pages [22].

5.1 Link probability versus lexical distance

An interesting regularity was discovered by projecting the distributional similarity data onto the content and link

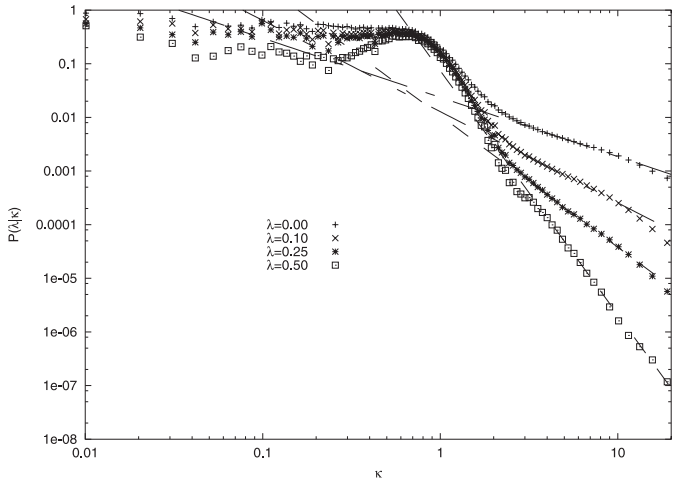

Fig. 7. Link probability versus lexical distance for Web pages based on the ODP sample. A nonlinear least-squares fit of the tail of each distribution to the power law model $\Pr(\lambda|\kappa) \sim \kappa^{-\gamma}$ is also shown. Data from [21].

similarity axes [21]. The idea was to quantify the dependence of link probability on content similarity (actually lexical distance, defined from TF-based cosine similarity via Eq. (1)). Since link probability is negligibly small and thus hard to measure in a large sparse network, I considered instead the conditional probability that the link similarity between two articles or pages is above some threshold λ, given the two documents have some lexical distance κ, as a function of κ:

$$\Pr(\lambda|\kappa) = \frac{|(p,q) : \delta_c(p,q) = \kappa \wedge \sigma_l(p,q) > \lambda|}{|(p,q) : \delta_c(p,q) = \kappa|} \qquad (20)$$

where p, q are two articles or Web pages. Figure 7 shows an interesting phase transition observed from the ODP sample of Web pages. There are two distinct regions around a critical distance κ^* independent of λ. For $\kappa < \kappa^*$ the probability that two documents are neighbors does not seem to depend on their lexical distance. For $\kappa > \kappa^*$ the probability decreases according to a power law $\Pr(\lambda|\kappa) \sim \kappa^{-\gamma}$, where the decay exponent γ grows linearly with λ ($\gamma \approx 6.4\lambda + 1$).

5.2 Similarity based growth models

The empirical power law tail of Figure 7 quantifies how the probability that two pages are linked decays with their content similarity. The same analysis, with similar results, was carried out for a collection of 15,785 articles published in the Proceedings of the National Academy of Sciences USA (PNAS) between 1997 and 2002 [22]. These results suggests that authors use content information when creating hyperlinks in Web pages, or citations in articles. Yet one does not find any reference to content in the recent literature on growth models for scale free networks, including the Web [3,23–26].

Most existing growth models are based on some form of *preferential attachment,* whereby one node at a time is added to the network with new edges to existing nodes selected according to some probability distribution. In the best known preferential attachment model a node i receives a new edge with probability proportional to its current degree, $\Pr(i) \propto k(i)$ [25]. This so-called BA model generates networks with power law degree distributions, in which the oldest nodes are those with highest degree. The *copying* model and its extensions implement equivalent rich-get-richer processes based on local walks, without requiring explicit knowledge of degree [27–29]. To give newer nodes a chance to compete for links, an extension of the preferential attachment model is based on linking to a node based on its degree with some probability or to a uniformly chosen node with the remaining probability [30,31]. Such a *mixture* model generates networks that can fit the power law degree distribution of the entire Web as well as the different distributions observed in subsets of the Web such as university and business homepages [32].

All the above models are capable of predicting the scale free degree distribution of Web pages and scientific articles, and the mixture model can predict non scale free distributions as well. However, none of those models can predict the distribution of lexical similarity across linked documents (Web pages connected by hyperlinks and documents connected by citations). Too see why, consider the distribution of lexical similarity across pairs of documents. If one counts all pairs, the distribution is roughly exponential: $\Pr(\sigma_c) \sim 10^{-\mu\sigma_c}$ where $\mu = 7$ for Web pages [21] and $\mu = 8$ for PNAS articles [22]. The distributions across linked documents, however, are qualitatively different. They have peaks at $\sigma_c > 0$ and decrease much more slowly for $\sigma_c \to 1$ [22]. One must conclude that content plays a role in the evolution of information networks. Put another way, if one simulates the growth models in the literature [25,27,32] using an exponential background distribution for σ_c, the distribution of σ_c across linked documents generated by the simulations is also exponential because σ_c is ignored by the models. This contradics the data, leading to the same conclusion.

A simple growth model that accounts for lexical similarity can be obtained by modifying the class of mixture models. This class has a free parameter that can be tuned to fit the data. At each step one new document is added and m new links or references are created from it to existing documents. At time t the probability that the ith document is selected and linked from the tth document is

$$\Pr(i) = \alpha \frac{k(i)}{mt} + (1-\alpha)\,\overline{\Pr}(i) \qquad (21)$$

where $i < t$ and $\alpha \in [0,1]$ is a preferential attachment parameter. In the classic mixture mixture model $\overline{\Pr}(i) = 1/t$, the uniform distribution [32]. Let us introduce an alternative *degree-similarity mixture* model in which

$$\overline{\Pr}(i) \propto [\delta_c(i,t)]^{-\gamma} = \left[\frac{1}{\sigma_c(i,t)} - 1\right]^{-\gamma} \qquad (22)$$

where γ is a constant. This model is inspired by the idea that authors tend to link new documents to popular and

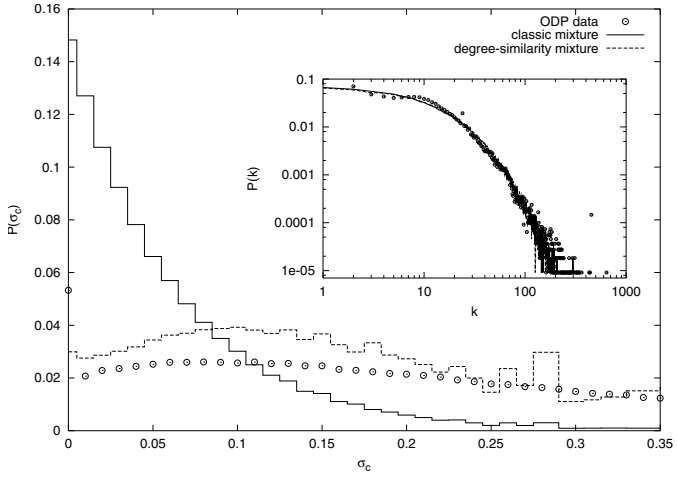

Fig. 8. Distribution of content similarity among linked Web pages and of degree (inset) predicted by simulating the two mixture models. In the classic mixture model simulation $\alpha = 0.3$, in the degree-similarity simulation $\alpha = 0.2$ and $\gamma = 1.7$. All parameters are set by matching or fitting the ODP data. Data from [22].

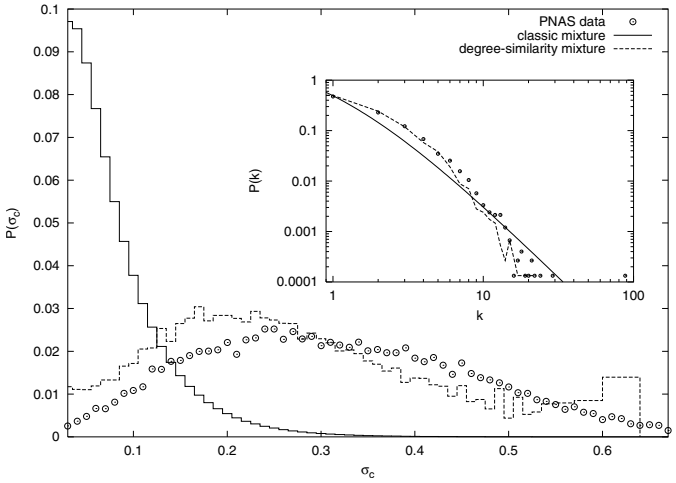

Fig. 9. Distribution of content similarity among titles and abstracts of articles that cite one another and of degree (inset) predicted by the two mixture models. In the classic mixture model simulation $\alpha = 0.5$, in the degree-similarity simulation $\alpha = 0.1$ and $\gamma = 3.1$. All parameters are set by matching or fitting the PNAS data (only references within the PNAS collection are considered). Data from [22].

related ones, and by the observation that link probability between two documents decays for large lexical distance as a power law $\Pr(\lambda = 0.1|\kappa) \sim \kappa^{-\gamma}$ where $\gamma = 3.1$ for PNAS articles [22] and $\gamma = 1.7$ for Web pages [21] (cf. Fig. 7). The free parameter α in the degree-similarity mixture allows to explicitly model the tradeoff between linking to related (similar) versus popular (high degree) documents.

5.3 Validation on Web and PNAS datasets

To validate the degree-similarity mixture model, the networks of Web pages and PNAS articles were built by simulation and compared to those obtained by simulating the classic mixture model. Figure 8 shows the predictions generated for Web pages. While both models accurately predict the degree distribution, only the degree-similarity mixture model reasonably approximates the similarity distribution of the ODP data.

The PNAS article data was analyzed analogously. Figure 9 shows the predictions generated by simulating the growth of the article network according to the two mixture models. Both models accurately predict the distribution of citation counts, although the degree-similarity model fits the PNAS data better. And again, the degree-similarity mixture model generates a similarity distribution in remarkable agreement with the data.

6 Conclusion

In this paper I reviewed a number of results that highlight the strong connections between different topologies in the Web and other document networks. These connections uncover a rich and complex relationship between the content of documents, their meaning, and the network structure that results from the links between documents created by authors.

The focus of different communities on different topologies (for example, lexical topology in information retrieval and link topology in statistical physics) may have hindered our progress in understanding the complex dynamics that govern document networks. For example, growth models based on just one topology are not realistic but their failure is not obvious unless one tests their ability to predict features related to different topologies. While search engine companies are trying to analyze different sources of evidence for identifying relevant documents, the scientific communities must also come together to gain new insight into the evolving structure of the Web and information networks. This may lead to more effective authoring guidelines as well as improved ranking, classification, clustering, and crawling algorithms.

The work reviewed here is currently being extended in a number of directions. As discussed in Section 2, a better semantic similarity measure is needed in order to take full advantage of the complex network ontologies provided by Web directories and classification schemes of digital libraries. We are currently studying a measure based on the maximum flow between two nodes, with edge capacities induced by node entropy.

It would be desirable to build a framework capable of efficiently computing correlations and maps based on arbitrary similarity measures. This way one could analyze and combine a large number of lexical and link similarity metrics to identify those that best approximate semantic relationships. The work outlined here is limited by its brute-force algorithm with quadratic complexity, which does not scale well with larger document collections.

The degree-similarity mixture model is being further validated by testing its ability to predict additional properties of the networks, such as clustering coefficient and degree correlation [6,29]. Finally, further insight must be gained by studying the relationship between the mechanism studied here (linking similar documents) and other processes likely to play a role in the the evolution of document networks, such as copying [29] and coauthorship [5,6].

I am grateful to Jon Kleinberg, Soumen Chakrabarti, Rob Axtell, László Barabási, Reka Albert, Mark Newman, Lada Adamic, Katy Börner, Padmini Srinivasan, Nick Street, and Alessandro Vespignani for helpful discussions on various aspects of the work reviewed in this paper. Thanks to the Open Directory Project for the ODP data, and to the National Academy of Sciences for the PNAS data. This work was funded by NSF Career Award No. IIS-0133124/0348940.

References

1. D. de Solla Price, Science **149**, 510 (1965)
2. R. Albert, A.-L. Barabási, Rev. Mod. Phys. **74**, 47 (2002)
3. S. Dorogovtsev, J. Mendes, *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford University Press, Oxford, UK, 2003)
4. R. Pastor-Satorras, A. Vespignani, *Evolution and Structure of the Internet* (Cambridge University Press, Cambridge, UK, 2004)
5. M. Newman, Proc. Natl. Acad. Sci. USA (2004)
6. K. Börner, J. Maru, R. Goldstone, Proc. Natl. Acad. Sci. USA (2004)
7. D. Wilkinson, B. Huberman, Proc. Natl. Acad. Sci. USA (2004)
8. P. Srinivasan, J. Amer. Soc. Inf. Sci. Techn. (forthcoming)
9. G. Salton, M. McGill, *An Introduction to Modern Information Retrieval* (McGraw-Hill, New York, NY, 1983)
10. C. Fox, *Information Retrieval: Data Structures and Algorithms* (Prentice-Hall, 1992)
11. M. Porter, Program **14**, 130 (1980)
12. P. Srinivasan, *Information Retrieval: Data Structures and Algorithms* (Prentice-Hall, 1992)
13. K. Sparck Jones, J. Documentation **28**, 111 (1972)
14. G. Salton, C. Buckley, Information Processing and Management **24**, 513 (1988)
15. S. Deerwester, S. Dumais, F. GW, T. Landauer, R. Harshman, J. Amer. Soc. Inf. Sci. **41**, 391 (1990)
16. *WordNet: An Electronic Lexical Database*, edited by C. Fellbaum (MIT Press, Cambridge, MA, 1998)
17. D. Lin, *Proc. 15th Intl. Conference on Machine Learning*, edited by J. Shavlik (Morgan Kaufmann, San Francisco, CA, 1998), pp. 296–304
18. F. Menczer, J. Amer. Soc. Inf. Sci. Technol. (2004) (forthcoming)
19. F. Menczer, *Poster Proc. 13th International World Wide Web Conference* (2004)
20. S. Brin, L. Page, Computer Networks **30**, 107 (1998)
21. F. Menczer, Proc. Natl. Acad. Sci. USA **99**, 14014 (2002)
22. F. Menczer, Proc. Natl. Acad. Sci. USA **101**, 5261 (2004)
23. R. Albert, H. Jeong, A.-L. Barabási, Nature **401**, 130 (1999)
24. B. Huberman, L. Adamic, Nature **401**, 131 (1999)
25. A.-L. Barabási, R. Albert, Science **286**, 509 (1999)
26. L. Adamic, B. Huberman, Science **287**, 2115 (2000)
27. J. Kleinberg, S. Kumar, P. Raghavan, S. Rajagopalan, A. Tomkins, Lecture Notes in Computer Science **1627**, 1 (1999)
28. S. Kumar et al., *Proc. 41st Annual IEEE Symposium on Foundations of Computer Science* (IEEE Computer Society Press, Silver Spring, MD, 2000), pp. 57–65
29. A. Vazquez, Phys. Rev. E **67**, 056104 (2003)
30. S. Dorogovtsev, J. Mendes, A. Samukhin, Phys. Rev. Lett. **85**, 4633 (2000)
31. C. Cooper, A. Frieze, *Proc. 9th Annual European Symposium on Algorithms*, edited by F. Meyer auf der Heide (Springer, Berlin, 2001), Vol. 2161 of Lecture Notes in Computer Science, pp. 500–511
32. D. Pennock, G. Flake, S. Lawrence, E. Glover, C. Giles, Proc. Natl. Acad. Sci. USA **99**, 5207 (2002)