Inferring Web Communities from Link Topology

David Gibson
Dept. of Computer Science
UC Berkeley
Berkeley, CA 94720 USA
+1 510 643 5425
dag@cs.berkeley.edu

Jon Kleinberg
Dept. of Computer Science
Cornell University
Ithaca, NY 14853 USA
+1 607 254 4636
kleinber@cs.cornell.edu

Prabhakar Raghavan
Almaden Research Center
IBM
San Jose, CA 95120 USA
+1 408 927 1804
pragh@almaden.ibm.com

ABSTRACT

The World Wide Web grows through a decentralized, almost anarchic process, and this has resulted in a large hyperlinked corpus without the kind of logical organization that can be built into more traditionally-created hypermedia. To extract meaningful structure under such circumstances, we develop a notion of *hyperlinked communities* on the WWW through an analysis of the link topology. By invoking a simple, mathematically clean method for defining and exposing the structure of these communities, we are able to derive a number of themes: The communities can be viewed as containing a core of central, "authoritative" pages linked together by "hub pages"; and they exhibit a natural type of hierarchical topic generalization that can be inferred directly from the pattern of linkage. Our investigation shows that although the process by which users of the Web create pages and links is very difficult to understand at a "local" level, it results in a much greater degree of orderly high-level structure than has typically been assumed.

Keywords: Hypertext communities, information exploration, World Wide Web, collaborative annotation.

INTRODUCTION

As hyperlinked environments grow in size and complexity, discovering and representing meaningful high-level structure in them becomes an increasingly challenging problem. This is particularly evident in the case of the World Wide Web (WWW), and the search for structure here is compelling for several reasons. The WWW is a hypertext corpus of enormous complexity, and it continues to expand at a phenomenal rate. Moreover, it can be viewed as an intricate form of "populist hypermedia," in which millions of on-line participants, many with conflicting goals, are continuously creating hyperlinked content. Thus, while individuals can impose order at an extremely local level, its global organization is utterly "unplanned" — in some sense, high-level structure can emerge only through *a posteriori* analysis.

The link structure of the WWW represents a considerable amount of latent human annotation, and thus offers a promising starting point for structural studies of the Web. There has been a growing amount of work directed at the integration of textual content and link information for the purpose of organizing [2, 4, 13, 24], visualizing [7, 21] and searching [1, 5, 6, 14, 17, 19, 22, 23, 25, 26] in hypermedia such as the WWW. The present work originates from the problem of searching on the WWW; building on this, it attempts to deal explicitly with defining a meaningful notion of *structure* in such an environment, as a way of addressing issues such as navigation and information discovery.

Our emphasis here is on an investigation of the link topology of the WWW, and some fairly pervasive themes we have identified about the structure of hypertextual *communities* that have developed in the Web. We will see that this notion of a *community* provides a surprisingly clear perspective from which to view the seemingly haphazard development of the Web's infrastructure.

The themes that emerge are valuable in a number of respects. Our analysis of the link structure of the WWW suggests that the on-going process of page creation and linkage, while very difficult to understand at a local level, results in structure that is considerably more orderly than is typically assumed. Thus, it gives us a global understanding of the ways in which independent users build connections to one another in hypermedia that arises in a distributed fashion, and it provides a basis for predicting the way in which on-line communities in less computer-oriented disciplines will develop as they become increasingly "wired." It also suggests some of the types of structured, higher-level information that designers of information discovery tools may be able to provide both for, and about, user populations on the Web.

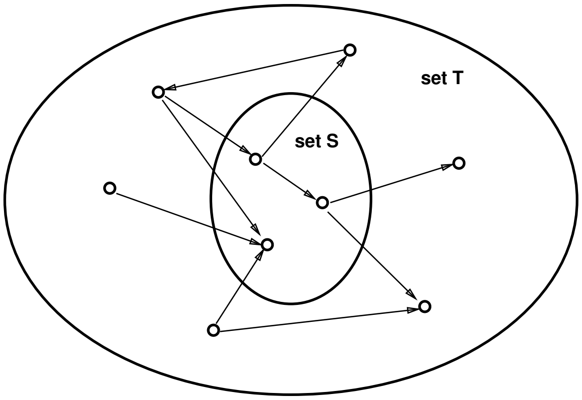
Our study is based on experience with a hyperlink-oriented method for searching introduced by Kleinberg in [17], and with HITS (Hyperlink-Induced Topic Search), an experimental system built around this technique. The underlying technique is discussed in detail in [17]. Here we invoke this technique, developing it for our study of communities on the Web. We begin with a brief summary (in the following section) of the main concepts from [17] that are necessary for understanding our study. Following this, the bulk of the paper is then a discussion of the methodology and basic motivation underlying our investigation of WWW communities, and the main themes that have emerged.

HITS is concerned with the identification of *authoritative* hypermedia sources for *broad-topic information discovery.* We briefly illustrate these notions with a set of examples. Consider, on the one hand, an individual who wishes to use the Web to find the phone number of a friend who is a student at Harvard Law School. The desired information is likely to be contained on only one or two pages, and hence the main issue is to locate this small number of *relevant* pages. On the other hand, consider an individual who wishes to use the Web to find information about Harvard University. Since there are more than $800,000$ pages on the WWW containing the term "Harvard," a shortage of relevant pages is no longer the problem. Rather, for a *broad-topic search* such as this, the user requires a way of identifying a small collection of the most central, or *authoritative*, pages on the topic "Harvard." Most standard search engines do not, for example, return authoritative pages such as `www.harvard.edu`. The interesting question arising in this context is: how can one determine, without human intervention, that `www.harvard.edu` is indeed a page that should be considered *authoritative* for the topic "Harvard"?

The technique underlying HITS stems from two premises [17]: first, that the implicit annotation provided by human creators of hyperlinks contains sufficient information to infer a notion of "authority"; and second, building on this, that sufficiently broad topics contain embedded *communities* of hyperlinked pages. We view such communities as containing two distinct, but interrelated, types of pages: *authorities* (highly-referenced pages) on the topic, as well as numerous pages that "point" to many of the authorities, and thus serve to "pull" them together. We refer to pages of the latter type as "hubs," since they serve as strong central points from which authority is "conferred" on relevant pages. Hubs and authorities exhibit what could be called a *mutually reinforcing relationship*: a good *hub* points to many good authorities; a good *authority* is pointed to by many good hubs. We break this apparent circularity by an iterative method described in the next section. An important point: our use of the term *community* does not imply that these structures have been constructed in a centralized or planned fashion. Rather, our experimentation with HITS shows that such communities of hubs and authorities *are a recurring consequence of the way in which creators of pages on the* WWW *link to one another in the context of topics of widespread interest.*

OVERVIEW OF HITS

We begin with a summary of the main concepts from [17] that are necessary for understanding the present work.



1. Starting from a user-supplied query, HITS assembles a *root set* S of pages: typically, up to 200 pages returned by a search engine such as AltaVista [9] on that query. It then expands this to a larger *base set* T by adding in any pages that point to, or are pointed to by, any page in S. (To prevent the size of T from exploding, only a fixed-size subset of the pages pointing to, or pointed to by, any single page in S is considered.) See the accompanying figure, in which we depict the relationship of the root set S and the base set T.

2. We associate with each page p a *hub weight* $h(p)$ and an *authority weight* $a(p)$, all initialized to 1. Let $p \to q$ denote "page p has a hyperlink to page q". HITS then iteratively updates the h's and a's as follows:

$$a(p) := \sum_{q \to p} h(q),$$
$$h(p) := \sum_{p \to q} a(q).$$

   Thus, a single iteration replaces $a(p)$ by the sum of the $h()$'s of pages pointing to p, and then replaces $h(p)$ by the sum of the $a()$'s of pages pointed to by p.

3. The updating operations are performed for all the pages, and the process is repeated (normalizing the weights after each iteration). It can be proved that this iterative process converges to stable sets of authority and hub weights [17]. We declare the 10 pages with the highest

$a()$ values together with the 10 pages with the highest $h()$ values to be the core of a community. (The number 10 here is more or less arbitrary, and is not crucial for the following discussion; essentially, we wish our communities to have a size that is "manageable" for human users.)

In fact it can be proved that the equilibrium values of the hub weights and authority weights correspond to coordinates in the principal eigenvectors of a pair of matrices M_{hub} and M_{auth} derived from the link structure [17].

Note that the method is extremely simple and mathematically clean: one can analyze its convergence properties in a rigorous fashion, and the only tunable parameter is the procedure for fixing the root set. We feel that this makes the technique an appealing framework in which to search for inherent structure in Web communities. The fact that the method is designed to run on an arbitrary link structure, without fine-tuning or the incorporation of expert knowledge about the WWW, suggests that the structural observations that emerge are largely intrinsic to the Web itself, rather than an artifact of an "over-trained" algorithm.

We return to our initial example, the topic "Harvard", to illustrate these notions in a concrete setting. The top authorities for this topic, as generated by HITS are the following. (The root set consisted of 200 pages returned by AltaVista on the query `"harvard"`.)

```
www.harvard.edu
www.hbs.harvard.edu
www.law.harvard.edu
ksgweb.harvard.edu
```

The top hubs in this case consist of pages created by various individuals not necessarily located at Harvard, consisting of links to a large number of the top authorities. Many other examples of a similar nature, for a range of topics, can be found in [17].

Note the crucial fact that the textual *content* of the pages involved is only considered in the initial step, when a *root set* is assembled from a search engine. Following this, the algorithm simply propagates weight over links without further regard to the *relevance* of the pages it is working with. The fact that HITS can reliably identify pages that are not only *authoritative* but also relevant to the user's initial query implies something about the *breadth* of the topic: since the initial root set was sufficiently rich in relevant pages, the densest community of hubs and authorities in the surrounding base set was relevant as well.

In general, of course, the base set contains not only this densest community but a large number of other meaningful communities as well [17]. For example, HITS also identifies, in the base set for the topic `"Harvard"`, a large community of pages on bio-medical topics, drawn into the base set because of the strong linkage between these pages and the many biological and medical labs associated with Harvard. This is a case in which the main community can be considered "on-topic," with several smaller "off-topic" communities arising as well.

It turns out that the matrices M_{hub} and M_{auth} discussed above contain enough information to allow for the discovery of such additional communities. Recall that the basic algorithm for identifying a community could be analyzed in terms of the *principal eigenvectors* of M_{hub} and M_{auth}. It turns out that the first few *non-principal eigenvectors* of M_{hub} and M_{auth} have the same intuitive meaning as the principal eigenvector: they represent pairs of weight assignments exhibiting the *mutually reinforcing relationship* of hubs and authorities [17]. Thus, by computing several of the non-principal eigenvectors of M_{hub} and M_{auth}, HITS can discover additional communities within the same *base set* of linked pages. We will use the term *principal community* to refer to the community found by the basic algorithm, associated with the principal eigenvectors of M_{hub} and M_{auth}; it is intuitively the community exhibiting the *densest* pattern of linkage between hubs and authorities. The additional communities associated with non-principal eigenvectors will be called *non-principal communities.* Since the non-principal eigenvectors of M_{hub} and M_{auth} can be naturally ordered (by the magnitude of their associated eigenvalues), this induces a natural ordering on the non-principal communities as well.

Multiple communities can also form in the base set because a query term has several meanings in different contexts; for example, the topic `"geometry"` produces communities on computational geometry, differential geometry, and seismic geometry.

RELATED WORK

The use of eigenvectors for the purposes of partitioning a graph was introduced by Donath and Hoffman in [10] and has been studied extensively since. The method underlying HITS is technically distinct from that of [10] (it does not partition the "Web graph"), though the heuristic intuition underlying both is clearly quite similar. For non-hyperlinked corpora, an information retrieval technique known as *latent semantic indexing* [8] makes use of the singular vectors of a matrix derived from the inverted index of the corpus. HITS, on the other hand, is operating purely on the link structure of a hyperlinked corpus, and makes no use of matrices with term weights.

The use of link information to improve search performance on the WWW has been advanced in previous work; hyperlink analysis has been used for enhancing relevance judgments [1, 12, 14, 19, 26], as well as for "ranking" WWW pages [5, 22, 23, 17]. Link structures have been studied in hypertext research that predates the WWW; in particular, Botafogo et al. [4] introduce graph-theoretic measures based on link density and node-to-node distances for clustering and searching in hypermedia. Their notions of *index* and *reference* nodes bear a relation to the notions of hubs and authorities used here; however, they are based purely on the out- and in-degrees of individual documents in the hyperlinked environment. (See [17] for a discussion of some of the difficulties in applying pure degree-counting methods to a domain on the scale of the WWW.)

The field of bibliometrics studies the patterns of citation — an implicit type of "linkage" — among scientific papers. See [27] for a review. A number of their measures have meaning in the context of hypermedia; some of these connections are studied in [18]. One can also interpret the behavior of HITS as relying on a type of *community memory*, as studied by Marshall et al. [20]. In essence, HITS is searching for a particular kind of link structure — reinforcing hubs and authorities — and this structure is created by communities of people collating useful information, whether for their own benefit or for others'.

METHODOLOGY AND OBSERVATIONS

We now discuss our use of HITS to study the emergence of communities on the Web; we also provide a high-level summary of our main observations. In the next section we discuss these observations in greater detail.

We focus on three issues:
(1) What are the (principal and non-principal) communities discovered by HITS?
(2) How do the communities discovered depend on the choice of root set? We report results on two kinds of variations of the root set: (i) assembling the root set from different sources (AltaVista vs InfoSeek), or from a query issued in different languages (`"astrophysics"` vs `"astrophysique"`); (ii) varying the size of the root set: we consider root sets composed of the top 25, 50, 100 and 200 results returned by AltaVista. This has the effect of "focusing" the text query to varying extents; we can then analyze the communities that result.
(3) How quickly do the communities crystallize as the number of iterations grows? (Recall that each iteration updates the hub and authority weights in terms of one another.) Are most communities tightly-knit (emerging after only a small number of iterations), or does it take many iterations for them to take shape? For instance after a single iteration, the top authorities are simply the pages in the base set with the largest number of incoming links. While these pages are indeed highly referenced (by definition), they tend to lack any thematic unity. Thus the top authorities for the topic `"Harvard"` after a single iteration consist of a mixture of pages for schools at Harvard, pages on bio-medical topics, and the home pages of YAHOO and Microsoft. As we increase the number of iterations, we are able to watch the base set "resolve" itself into coherent communities of hubs and authorities.

Issues (1) and (2) are discussed in detail below. To study issue (3) quantitatively, we define the *principal community*, as before, to be the set C of the top 10 authorities and top 10 hubs — 20 pages in all. Recall that this set C is implicitly a function of one parameter R, the root set size. However to study the "convergence" to C, we include N, the number of iterations, as second parameter; so we say that $C(R, N)$ is the principal community obtained by running HITS for N iterations starting from a root set size of R. In our experience with several hundred HITS trials on a variety of topics, the communities that form predictably become stable with a root set size of 200 and after 50 iterations; so we define $C^* = C(200, 50)$. Now, for various values of R and N, we can look at the *overlap* between $C(R, N)$ and C^* — the number of pages they have in common. In Figures 1–6, on the final page, we plot these overlaps for $N = 1, 3, 10, 50$ and $R = 25, 50, 100, 200$, for six representative topics in our study:

```
Harvard
cryptography
English literature
skiing
optimization
operations research
```

(Each curve in these plots measures overlap with respect to root set size, for a fixed value of N. Again, $N = 50$ essentially represents convergence in our case.)

Note that our representative topics have different levels of connection to the broad area of computer science, ranging from topics that are heavily computer-oriented (`"cryptography"`) to entirely different academic disciplines (`"English literature"`). Such topics also exhibit a range of different patterns and quantity of linkage, and this has an effect on the *rate* at which the principal community "crystallizes," in terms of the number of iterations and the root set size. These contrasts will be discussed further below. The plots in In Figures 1–6 also illustrate clearly the danger of using only a single iteration.

We now summarize some principal themes that emerge from our analysis. We find that communities on sufficiently broad topics tend to have a fairly "robust" structure: the groupings of pages discovered tend to be relatively independent of the exact choice of root set. We also find that the success of HITS depends on both the *breadth* of a topic and the discipline of human knowledge under which it falls. This is because the density and comprehensiveness of hyperlinking is far greater in some disciplines (e.g., computer science) than in others (e.g., subjects in the academic humanities, which are moving onto the Web more slowly.) On topics which do not have a sufficient density of hyperlinking, HITS tends to find communities that *generalize* the initial topic in one of several senses; this will be another focus of our discussion below.

Note that a fairly counter-intuitive point has been emerging from the development above. Specifically: *The greatest degree of orderly structure, as extracted by* HITS*, is found in communities for which the number of relevant pages, and the density of hyperlinking, is the largest.* We have seen this phenomenon with `"cryptography"` and `"English literature"`. This is in contrast to the standard point of view that the WWW is becoming increasingly "chaotic" and difficult to model; it suggests that the technique underlying HITS is actually becoming more effective as the size of the Web continues to increase. A consequence of this is that we can use our experience with highly-linked topics (e.g. `"cryptography"`) to make predictive statements about the structure of communities that have yet to fully "catch up" in the setting of the WWW.

The power of HITS on highly-linked communities is akin to a "law of large numbers": many independent, largely random annotations will reinforce one another and metamorphose into a broad-topic community replete with structure. We note that the emergence of regular structure in random networks is an active topic of research in combinatorics (see e.g. [3]), and we feel that it would be interesting to investigate such connections further in the context of HITS and the WWW.

THE STRUCTURE OF COMMUNITIES

We now detail our main findings on the structure of communities.

Robustness. For broad topics, HITS produces stable, robust communities despite starting from a very small sample of relevant pages in the initial root set. We have explored this by several direct methods, providing HITS with a variety of different root sets relevant to the same topic. For example, we issue the same query string to multiple search engines (e.g. AltaVista [9], Infoseek [16], and Excite [11]); this typically produces root sets with very little intersection. Similarly, one can obtain root sets that are nearly disjoint by issuing a query term in several different languages (e.g., `"astrophysics"` vs `"astrophysik"` vs `"astrophysique"`).

We find that the main communities tend to recur in all these experiments, regardless of how the root set is constructed. However, since communities have more or less representation in different base sets, the identity of the *principal* community is not always the same. But what this suggests is that multiple experiments with different root sets are providing us with multiple, slightly altered views of a small set of underlying communities, which are being sampled by the various choices of root sets.

To illustrate this point, consider the way in which the principal authorities for the topic `"astrophysics"` recur in non-principal communities for the French and German versions of the topic. The top 5 authorities for `"astrophysics"` are

```
fits.cv.nrao.edu/www/astronomy.html
cdsweb.u-strasbg.fr/Simbad.html
www.aas.org/
heasarc.gsfc.nasa.gov
adsabs.harvard.edu/abstract-service.html
```

For `"astrophysique"`, the top 5 authorities in the 8th non-principal community are

```
cdsweb.u-strasbg.fr/CDS.html
adswww.harvard.edu/
cdsweb.u-strasbg.fr/Simbad.html
adsabs.harvard.edu/abstract-service.html
fits.cv.nrao.edu/www/astronomy.html
```

For `"astrophysik"`, the top 5 authorities in the 7th non-principal community are

```
adswww.harvard.edu/
cdsweb.u-strasbg.fr/Simbad.html
adsabs.harvard.edu/abstract-service.html
aibn55.astro.uni-bonn.de:8000/
www.univ-rennes1.fr/ASTRO/astro.english.html
```

Topic Generalization. There is no sharp boundary between those topics that are "broad" and those that aren't; but one of the primary themes that emerges from our experience is that HITS tends to "generalize" topics that are not sufficiently broad. By this we mean that the principal community of hubs and authorities will be relevant to a topic which includes, but is larger than, the initial topic provided to HITS.

To make this concrete, we focus on three basic examples. Consider first the notion of topics defined by proper names. The topic `"Michael Jordan"` (the basketball star) turns out to be sufficiently broad: there are numerous hub pages containing links to pages on Michael Jordan and his team, and hence the principal community is highly relevant. On the other hand, the topic

"Dennis Ritchie" (an author of the C programming language) produces the following top 3 authorities:

www.cm.cf.ac.uk/Dave/C/CE.html
www.cyberdiem.com/vin/learn.html
www.lysator.liu.se/c/index.html

All of these are highly-referenced pages on the C programming language itself. Thus, while the principal community is relevant to a *generalization* of the topic "Dennis Ritchie," the individual himself has been swallowed up by the subject to which he most prominently belongs.

One sees a sense in which the mechanism of HITS pushes specific topics "upwards" in an implicit topic hierarchy, while sufficiently broad topics represent fixed points on which HITS finds relevant authorities that are not overly general. Consider, for example, the topic "optimization". The top three authorities are the home pages of the Institute for Operations Research and the Management Sciences, the CMU Graduate School of Industrial Administration, and the Society for Industrial and Applied Mathematics.

www.informs.org
mat.gsia.cmu.edu
www.siam.org

An examination of these and other authorities might suggest that the community is in fact relevant to a larger topic that includes much of the field of optimization — namely, operations research. This possibility is supported by finding the principal community for the topic "operations research"; the pages discovered have significant overlap with those for "optimization". The top three authorities are

www.informs.org
mat.gsia.cmu.edu
www.gams.com

Finally, it is interesting that topics which would naturally be considered "parallel" can behave differently with respect to this type of generalization. This can be due to differences in the amount of hyperlinking among the relevant pages. For example, the top authorities for "English literature" remain focused:

the-tech.mit.edu/Shakespeare/works.html
www.english.upenn.edu/~jlynch/Lit
humanitas.ucsb.edu/shuttle/eng-mod.html

The top authorities for "German literature", on the other hand, are relevant to a range of topics in European literature more generally.

www.crs4.it/HTML/Literature.html
humanities.uchicago.edu/ARTFL/ARTFL.html
un2sg1.unige.ch/www/athena/html/francaut.html

At the highest level, an examination of the two link structures indicates that the *density* of linkage is much greater for pages on the former topic than for those on the latter, and this leads to their different behavior. We shall have more to say about this contrast below.

Generalization is an interesting feature of HITS, since it allows for the automatic characterization of certain specific topics in terms of their generalizations. For example, it was purely through an analysis of the link structure that HITS placed the topic "Dennis Ritchie" in a community on the C programming language.

Convergent Generalization and a "Tree of Topics." The implicit topic hierarchy just discussed can be explored more fully through further experimentation with HITS. It is natural to picture the process of abstraction and generalization as occurring on a *tree of topics*: the most general topics (e.g. Science, Art, Recreation) are closest to the root, and their descendants represent sub-topics. Such an idea has been realized on the WWW, by human ontologists, through the construction of *searchable hierarchies* such as YAHOO. A searchable hierarchy includes hand-annotation of the various topics (e.g. Science and Art); however, even through a purely automated analysis of the links, we can gain information about such trees of topics.

The hypothesis of an underlying tree-based explanation for generalization is supported by the following phenomenon of *convergent generalization*, which HITS exhibits. We have found that, given a broad topic on which the technique does not generalize, one often discovers very similar communities by applying HITS to a range of more specific sub-topics.

One clear example is provided by the topic "cryptography". This topic is sufficiently general that HITS reliably finds very focused authorities and hub pages. Now, suppose we choose more specific sub-topics of cryptography — specifically, names of individual cryptographers. One finds that most of these sub-topics are generalized in very similar ways, to communities that have significant overlap both with each other and with the larger topic "cryptography".

Intuitively, we are looking at a portion of the topic tree in the vicinity of the node for "cryptography"; and we find that many of the child nodes generalize upward to this common parent.

Other Factors Affecting Generalization. In addition to the natural notion of generalization discussed above, we have identified a number of other factors that cause HITS to converge to communities that are not completely focused on the initial topic. These factors recur in a number of different contexts, and again seem to highlight some fundamental themes about the structure of hyperlinked communities on the Web, and the underlying interests of user populations in this domain. In particular, we will discuss the following two factors: "Web-centric" sub-topics, and commercialization.

The first of these is the notion that, ultimately, what determines the "generality" of a topic in this setting is its *representation* on the WWW. Thus, certain topics can seem artificially broad, and others artificially narrow, because they are of more or less interest to creators of Web pages. This provides another perspective from which to view the notion that HITS tends to be most effective within topics that have the most "wired" communities.

The simplest manifestation of this principle is the following: in a large fraction of cases, the topic that Web pages are most concerned with involves the Web itself; and this can influence, more or less subtly, the structure of the communities that HITS discovers. In particular, the principal community can at first appear to be a *specialization*, rather than a generalization, of the initial topic; but in reality, HITS has focused on this community because it represents a more "Web-centric" version of the topic.

We consider three examples. For the topic `"linguistics"`, the top authorities are

```
www.cs.columbia.edu/~acl/home.html
www.cs.columbia.edu/~radev/cgi-bin/universe.cgi
www.ling.rochester.edu/linglinks.html
```

The first two of these are strong authorities for the field of *computational linguistics*. While this is a sub-topic of the initial query, HITS has converged to it because of the same influences that typically cause it to generalize: in the setting of the WWW it is a topic with a considerably greater density of linkage than classical linguistics. Next, we consider the topic `"Harvard"` as of January 1997. Included among the top authorities are home pages for the Harvard Conference on the Internet and Society, a very large Internet conference held in May 1996. From the perspective of pages on the WWW, this is a prominent topic closely related to the initial query, `"Harvard"`. Finally, an observation which should be obvious in hindsight: The top authority for the topic `"physics"` is CERN, the location where the Web was in a sense "born":

```
www.cern.ch/
aps.org/
pdg.lbl.gov/
```

One can witness this phenomenon at work in the contrast between `"English literature"` and `"German literature"`. In particular, the greater link density of the former topic can be partially explained by a number of influences, including the increased representation of North American university home pages, which tend to be highly linked, and the fact that the creation of Shakespeare repositories is a much more developed enterprise on the WWW than is the comparable activity for German authors. In the other direction, we find it quite interesting that a large community associated with the output of `"German literature"` was in fact focused around Scandinavian literature; this relates closely to the numerous anecdotal observations about the unusually high density of linking among Web pages from Finland, Sweden, and Norway.

One way to view this general issue (though it is an oversimplification): at the root of our conceptual tree of topics sits the World Wide Web itself.

The other main influence we touch on here is that of commercial/advertising influences on the link structure. The process of page creation on the WWW has many simultaneous participants, and some of these can bring many resources to bear on engineering the link structure in a way that favors them. Thus, for topics with both commercial and individual involvement, the authorities in the principal community will overwhelmingly tend to be highly-commercialized pages. A simple example is the topic `"tennis"`:

```
www.tennisserver.com/
www.uspta.org/
espnet.sportszone.com/ten/
```

However, we stress that it is considerably harder to mislead HITS by "spamming" than to mislead a term-based search engine: HITS measures the collective authority vested in a resource by many people; in contrast, it is common to for creators of popular pages to try "boosting" their ranking on term-based engines by including large numbers of judiciously-chosen keywords. (As a pragmatic matter, one way to limit the self-conferral of authority is to limit – or even ignore – the propagation of weights along links into a page from pages in its own internet sub-domain.)

A phenomenon which combines both these influences, and which shows up quite frequently, is the recurring presence of pages such as AltaVista, YAHOO, and `www.microsoft.com` within a community of hubs and authorities, regardless of the topic. The point is that these pages are linked to so heavily, from essentially all portions of the WWW, that they frequently show up in lists of authorities. (Thus, for example, YAHOO was the ninth-ranked authority for the topic `"physics"`.) We mention this issue because it tells us something about the extent to which such pages have infiltrated hyperlinked communities in all disciplines; as a pragmatic issue, they could be removed from the output of HITS by essentially treating them as the WWW-equivalent of "stop words" in information retrieval.

Temporal Issues. Comparing the principal authorities for `"Harvard"`, August 1997, and `"Harvard"`, January 1997, brings up an additional theme concerning hyperlinked communities. There can often be substantial short-term factors that temporarily influence the set of principal authorities for a topic; in the above example, the Harvard Conference on the Internet and Society was still a prominent feature of the topic "Harvard" in January 1997, but not in August 1997. Another example is the topic `"comets"`; in May 1997 HITS discovered a large community concerned with alien visits and the "Heaven's Gate" group.

Short-term influences die out as pages and links are removed over time; though they can be artificially kept "current" by their inclusion in the indices of search engines. An interesting method for obtaining the long-term "core" of a topic is to superimpose the results of HITS on the same topic, spaced out over several-month periods.

Dissecting the Influences on a Topic. Earlier, we indicated that the technique underlying HITS, through the use of eigenvectors, is able to discover multiple communities associated with a given topic: a single *principal* community, together with an arbitrary number of sparser *non-principal* communities. We find that typically the influences leading to topic generalization, in the forms discussed above, are concentrated in some but not all of the communities discovered by HITS; and that some of the *non-principal* communities are more "purely" focused on the search topic. By examining all of the strongest communities together, one assembles a partially nested, partially overlapping arrangement of the different hyperlinked sub-communities that make up the search topic. For example, in the topic `"physics"`, one finds strong communities whose top authorities are composed entirely of (i) academic departments, (ii) high-energy accelerators and colliders, and (iii) professional societies. For the topic `"Harvard"` (Aug. 1997), the group of bio-medical pages can likewise be easily identified as a coherent non-principal community, separate from the home pages of schools at Harvard.

This type of dissection can also be an effective way to separate out the "Web-centric" influences on a topic. An example from [17] shows this very cleanly: starting from a root set of pages in the 2-step vicinity of `www.nytimes.com`, the principal authorities consist of a mixture of on-line news organizations and popular Internet sites. The first two non-principal communities separate this mixture very sharply:

```
Authorities in first non-principal community:
 www.microsoft.com/
 www.ibm.com/
 www.apple.com/
 www.hp.com/
 www.sun.com/

Authorities in second non-principal community:
 www.nytimes.com/
 www.usatoday.com/
 www.cnn.com/
 www.sjmercury.com/
 www.chicago.tribune.com/
```

FURTHER DIRECTIONS

The dynamic growth of the Web immediately suggests that the type of analysis we have performed should be repeated over time and the results compared. Such an approach provides a way of studying the temporal evolution of communities on the Web, and of understanding how the techniques considered here will adapt as the Web continues to grow in both size and complexity.

We are currently investigating ways of using HITS to improve performance on information retrieval tasks by combining text and the structure of hyperlinks; for work in this direction, see [6]. One goal here is to create a client-side information discovery system whose search parameters — at both the text-based and link-based levels — are fully "tunable" by individual users,

CONCLUSION

The WWW has grown into a hypertext environment of enormous complexity; and the *process* underlying its growth has been driven in a chaotic fashion by the individual actions of numerous participants. Our experience with HITS suggests, however, that in many respects the end product is not as chaotic as one might suppose: the *aggregate* behavior of user populations on the WWW can be studied through a mathematically clean technique for analyzing the Web's link topology, and one can use this technique to identify themes about *hyperlinked communities* that appear to span a wide range of interests and disciplines.

ACKNOWLEDGEMENTS

We have benefited from discussions with many people on the topics covered here. In particular, we thank Rakesh Agrawal, Soumen Chakrabarti, Byron Dom, Nimrod Megiddo, Christos Papadimitriou, Sridhar Rajagopalan, Ted Selker and Eli Upfal. The work of the first and second authors was performed in part while visiting the IBM Almaden Research Center. The second author is supported in part by an Alfred P. Sloan Research Fellowship and by NSF Faculty Early Career Development Award CCR-9701399.

REFERENCES

1. G.O. Arocena, A.O. Mendelzon, G.A. Mihaila, "Applications of a Web query language," *Proc. 6th International World Wide Web Conference*, 1997.
2. M. Q Wang Baldonado, T. Winograd, "SenseMaker: An information-exploration interface supporting the contextual evaluation of a user's interests," *Proc. ACM SIGCHI Conference on Human Factors in Computing*, 1997.
3. B. Bollobás, *Random Graphs*, Academic Press, 1985.
4. R. Botafogo, E. Rivlin, B. Shneiderman, "Structural analysis of hypertext: Identifying hierarchies and useful metrics," *ACM Trans. Inf. Sys.*, 10(1992), pp. 142–180.
5. J. Carrière, R. Kazman, "WebQuery: Searching and visualizing the Web through connectivity," *Proc. 6th International World Wide Web Conference*, 1997.
6. S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, S. Rajagopalan. "Automatic resource list compilation by analyzing hyperlink structure and associated text." *Proc. 7th International World Wide Web Conference*, 1998.
7. C. Chen. Structuring and visualizing the WWW by generalised similarity analysis. *Proc. 8th ACM Conference on Hypertext*, 177–186, 1997.
8. S. Deerwester, S. Dumais, T. Landauer, G. Furnas, R. Harshman, "Indexing by latent semantic analysis," *J. American Soc. Info. Sci.*, 41(1990).
9. Digital Equipment Corporation, *AltaVista search engine*, `altavista.digital.com/`.
10. W.E. Donath, A.J. Hoffman, "Algorithms for partitioning of graphs and computer logic based on eigenvectors of connections matrices," *IBM Technical Disclosure Bulletin*, 15(1972).
11. Excite Inc., *Excite*, `www.excite.com`.
12. M.E. Frisse, "Searching for information in a hypertext medical handbook," *Communications of the ACM*, 31(7), pp. 880–886.
13. R. Furuta, F. M. Shipman III, C. C. Marshall, D. Brenner and H-W. Hsieh. Hypertext paths and the world-wide web: experiences with Walden's paths. *Proc. 8th ACM Conference on Hypertext*, 167–176, 1997.
14. G. Golovchinsky. What the query told the link: the integration of Hypertext and Information Retrieval. *Proc. 8th ACM Conference on Hypertext*, 67–74, 1997.
15. G. Golub, C.F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, 1989.
16. Infoseek Corporation, *Infoseek search engine*, `www.infoseek.com`.
17. J. Kleinberg, "Authoritative sources in a hyperlinked environment," *Proc. ACM-SIAM Symposium on Discrete Algorithms*, 1998. Also appears as IBM Research Report RJ 10076(91892) May 1997, and at `www.cs.cornell.edu/home/kleinber/`.
18. R. Larson, "Bibliometrics of the World Wide Web: An exploratory analysis of the intellectual structure of cyberspace," *Ann. Meeting of the American Soc. Info. Sci.*, 1996.
19. M. Marchiori, "The quest for correct information on the Web: Hyper search engines," *Proc. 6th International World Wide Web Conference*, 1997.
20. C. C. Marshall, F. M. Shipman III, R. J. McCall, "Putting Digital Libraries to Work: Issues from Experience with Community Memories", *Proc. First Annual Conference on the Theory and Practice of Digital Libraries*, 1994.
21. S. Mukherjea and Y. Hara. Focus+Context Views of World-Wide Web Nodes. *Proc. 8th ACM Conference on Hypertext*, 187–196, 1997.
22. L. Page, "PageRank: Bringing order to the Web," Stanford Digital Libraries working paper 1997-0072.
23. L. Page, S. Brin, R. Motwani, T. Winograd, "The PageRank citation ranking: Bringing order to the Web," submitted for publication.
24. P. Pirolli, J. Pitkow, R. Rao, "Silk from a sow's ear: Extracting usable structures from the Web," *Proc. ACM SIGCHI Conference on Human Factors in Computing*, 1996.
25. E. Spertus, "ParaSite: Mining structural information on the Web," *Proc. 6th International World Wide Web Conference*, 1997.
26. R. Weiss, B. Velez, M. Sheldon, C. Nemprempre, P. Szilagyi, D.K. Gifford, "HyPursuit: A Hierarchical Network Search Engine that Exploits Content-Link Hypertext Clustering," *Proceedings of the Seventh ACM Conference on Hypertext*, 1996.
27. H.D. White, K.W. McCain, "Bibliometrics," in *Ann. Rev. Info. Sci. and Technology*, Elsevier, 1989, pp. 119-186.
28. Yahoo! Corp. *Yahoo!*, `www.yahoo.com`.

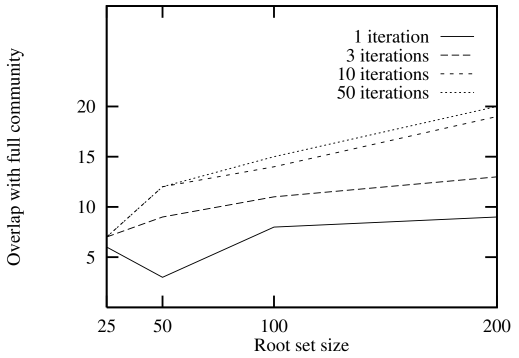


Figure 1. ‘‘Harvard’’

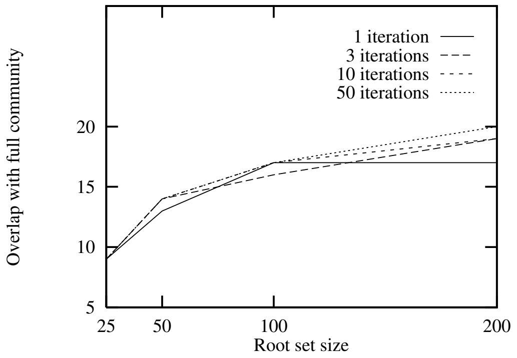


Figure 2. ‘‘Cryptography’’

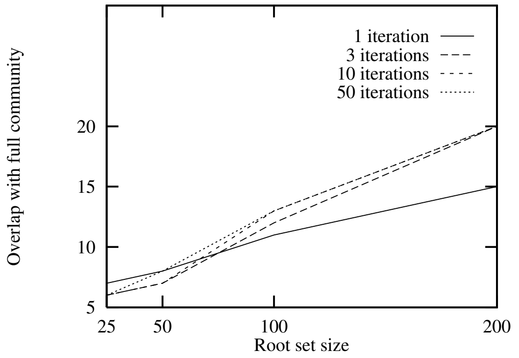


Figure 3. ‘‘English Literature’’

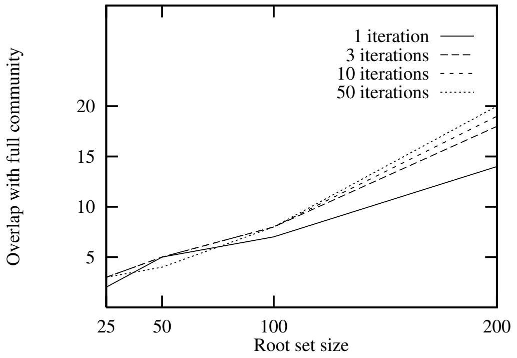


Figure 4. ‘‘Skiing’’

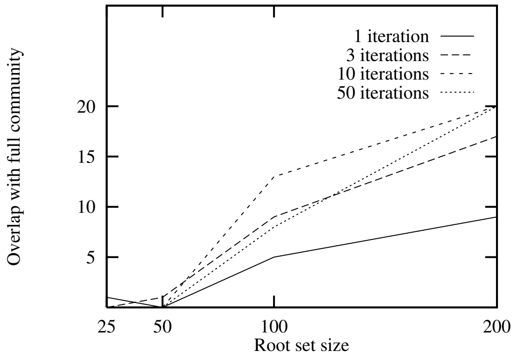


Figure 5. ‘‘Optimization’’

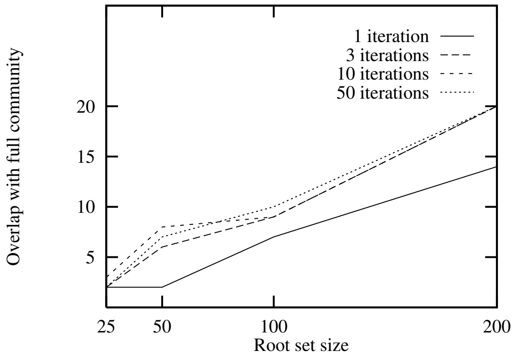


Figure 6. ‘‘Operations Research’’