5

Link Analysis

Computing the relevance of a document is a major issue in Web-based information retrieval. As we have seen in Chapter 4, when considering unstructured collections of documents, it is possible to compute relevance with respect to user queries such as those involving keyword-based Boolean expressions, or those involving measures of similarity between documents. In each of these cases the score is a function of the document and the query. Large collections of hypertext such as the Web are more interesting in this respect since they allow us to compute scoring functions that include topological information about the hypertext graph. The basic assumption is that hyperlinks contain information about the human judgment of a document. To a first approximation, the more incoming links that exist for a document, the more likely it is that the document was judged to be 'important' by the authors of other documents linking to it.

Incoming links embody a common notion of popularity that also exists in other domains or in other webs. Networks of interaction have been studied for a long time in social sciences (Wasserman and Faust 1994), where nodes correspond to persons or organizations, and edges represent some type of social interaction. Intuitively, increasing the number of incoming links to a node should increase a common-sense measure of standing or popularity or prestige for that node. However, it should be also common sense that just counting the number of links does not necessarily provide an accurate measure of standing. For example, measuring the prestige of an enterprise by the mere number of its clients could be misleading, since different clients may have very different weights.

An important example is the network obtained by considering the scientific literature. Nodes in this case are papers, books, or entire journals, and edges correspond to citations. It makes sense to assume that the more citations a paper or a book receives, the more it can be assumed to be important, since it has been judged as useful by other scientists. The systematic construction of such networks through citation indexes was introduced by Garfield (1955), who later proposed a measure of standing for journals that is still in use. This measure, called *impact factor* (Garfield 1972), is defined as the average number of citations per recently published item. More precisely, if C is the total number of citations in a given time interval $[t, t + t_1]$ to articles published

Modeling the Internet and the Web P. Baldi, P. Frasconi and P. Smyth

ISBN: 0-470-84906-1

by a given journal during $[t - t_2, t]$, and N is the total number of articles published by that journal in $[t - t_2, t]$, the impact factor is defined as C/N (typically $t_1 = 1$ year and $t_2 = 2$ years). Thus, the impact factor is a very simple measure, since it basically corresponds to the normalized indegree of a journal in a subgraph of the citation network. Graph-based link analysis for the scientific literature goes back to the 1960s (Garner 1967). However, these ideas were not exploited in the development of first-generation Web search tools.

The paper by Bray (1996) reports an early attempt to apply social networks concepts to the Web. He suggested a Web visualization approach where the '... appearance of a site should reflect its *visibility*, as measured by the number of other sites that have pointers to it ...' and '... its *luminosity*, as measured by the number of pointers with which it casts navigational light off-site...'. Visibility and luminosity defined in this way are directly related to the indegree and the outdegree of websites, respectively. More recently, toward the end of the 1990s, link analysis methods became more widely known and used in a search engine context, leading to what is sometimes called the *second generation* of Web searching tools.

This chapter reviews the most common approaches to link analysis and how these techniques are be applied to compute the popularity of a document or a site. The algorithms presented in this chapter extract emergent properties from a complex network of interconnections, attempting to model (indirectly) subjective human judgments. It remains debatable as to whether popularity (as implied by the mechanism of citations) captures well the notions of relevance and quality as they are subjectively perceived by humans, and whether link analysis algorithms can successfully model human judgments.

5.1 Early Approaches to Link Analysis

Our notation for hypertext will be straightforward. For each vertex v in the hypertext graph $G = (V, E)$, $d(v)$ denotes the contents of the document at vertex v. If $d(v)$ is considered to be an isolated document, then its score with respect to a query q is $s(v \mid d(v), q)$. When considering $d(v)$ in its hypertext context, the score should also depend on G and will be denoted as $S(v \mid d(v), q, G)$. In the following, we will simplify the notation of these scores by just writing $s(v)$ and $S(v)$ if the dependencies on $d(v)$, q, and G are obvious from the context.

To quantify visibility (luminosity) $S(v)$ could simply be designed to grow with the indegree (outdegree) of v as hinted by Bray (1996). Clearly, however, such an approach suffers from a fundamental limitation: it would fail to capture the relative importance of different parents (children) in the graph. For example, a Web page with a small number of links coming from important sites should be considered more popular than a Web page with a larger number of links and whose sources are all from unimportant or irrelevant sites. Hence, rather than a mere count, popularity should be computed as a weighted sum of the citations a document receives through hypertext links. Ideas having this flavor are less recent than we might expect.

The use of hypertext information in information retrieval is older than the Web. Mark (1988), for example, was concerned with retrieving hypertext cards in a medical domain and noted that '... often cards do not even mention what they are about, but assume that the reader understands the context because he or she has read *earlier* cards.' He then proposed a simple algorithm for scoring documents where relevance information was transmitted from documents to their parents in the hypertext graph G. More precisely, the 'global' score of v given the query and the topology of G was computed as:

$$S(v) = s(v) + \frac{1}{|\,\mathrm{ch}[v]|} \sum_{w \in |\,\mathrm{ch}[v]|} S(w). \tag{5.1}$$

This simple algorithm somewhat resembles message passing schemes that are very common in connectionism (McClelland and Rumelhart 1986) or in graphical modeling (Pearl 1988). As such, it requires G to be a DAG so that a topological sort[1] can be chosen for updating the global scores S. The DAG assumption is reasonable in small hypertexts with a root document and a relatively strong hierarchical structure (in this case, even if G is not acyclic, not much information would be lost by replacing it with its spanning tree). The Web, however, is a large and complex graph. This may explain why search engines largely ignored its topology for several years.

The paper by Marchiori (1997) was probably the first one to discuss the quantitative concept of *hyper information* to complement *textual information* in order to obtain the *overall information* contained in a Web document. The idea somewhat resembles Frisse's approach. Indeed, if we rewrite Equation (5.1) as

$$S(v) = s(v) + h(v), \tag{5.2}$$

then $s(v)$ can be thought of as the textual information (that only depends on the document and the query), $h(v)$ corresponds to the hyper information that depends on the link structure where v is embedded, and $S(v)$ is the overall information. Marchiori (1997) did not cite Mark (1988), but nonetheless he identified a fundamental problem with Equation (5.1). If an irrelevant page v has a single link to a relevant page w, Equation (5.1) implies that $S(v) \geqslant S(w)$. The scenario would be even worse in a chain of documents $v_0, v_1, \dots, v_k$. Here if $S(v_k)$ is very high but $S(v_0), \dots, S(v_{k-1})$ are almost zero, then v_0 would receive a global score higher than v_k, even though a user would need k clicks to reach the important document.

As a remedy, Marchiori suggested that in this case the hyper information of v_0 should be computed as

$$h(v) = \sum_{w \in \mathrm{ch}[v]} F^{r(v,w)} S(w), \tag{5.3}$$

where $F \in (0, 1)$ is a fading constant and $r(v, w) \in \{1, \dots, |\mathrm{ch}[v]|\}$ is the rank of w after sorting (in ascending order) the children of v according to the value of

[1] A topological sort is an ordering '$<$' of the vertices such that $v < v'$ if and only if there is a directed path from v' to v (Cormen *et al.* 2001).

$S(w)$. When applying Equation (5.3) to a linear chain $v_0, v_1, \ldots, v_k$ of documents, one would get $S(v_0) = \sum_{i=1}^{k} F^i S(v_i)$, so the contribution of the score of v_k fades exponentially as one moves back in the graph. As it turns out, Equation (5.3) in general implies a recursive form of computation that cannot be carried out in a cyclic graph. Marchiori (1997) suggested a solution to this problem assuming a finite horizon of propagation, i.e. $S(v)$ was computed on the tree rooted at v and having a fixed small depth k.

Before we discuss link analysis of the Web graph, it will first be useful to establish certain basic mathematical results relating to nonnegative matrices, graphs, and Markov chains. The reader familiar with these topics can safely skip the next section.

5.2 Nonnegative Matrices and Dominant Eigenvectors

A square matrix $\boldsymbol{A}$ is said to be *nonnegative*, written $\boldsymbol{A} \geqslant 0$, if all its elements are nonnegative. Important examples of nonnegative matrices include graph incidence (of adjacency) matrices and stochastic matrices. Given a directed graph $G = (V, E)$, the incidence matrix $\boldsymbol{A}$ of G is defined as the 0-1 matrix with $a_{ij} = 1$ if and only if $(i, j) \in E$. Note that for simplicity we have assumed that the vertices in V are identified by the integers $1, 2, \ldots, n$, where $n = |V|$. Stochastic matrices are commonly used to describe first-order Markov chains as discussed in Appendix A. In a system with n discrete states, each entry of a stochastic matrix $\boldsymbol{A}$ contains the transition probability $a_{ij} = P(S_t = j \mid S_{t-1} = i)$. In this case, the probability axioms imply that each $a_{ij} \geqslant 0$ and that the elements of each row i should sum to unity.

A nonnegative $n \times n$ matrix $\boldsymbol{A}$ is said to be *irreducible* if, for each pair of indices (vertices) i and j, there exists a corresponding integer t such that $(\boldsymbol{A}^t)_{ij} > 0$. If $\boldsymbol{A}$ is the adjacency matrix of a (directed) graph, this property tells us that the graph is (strongly) connected. By contrast, a reducible matrix is associated with a graph with more than one (strongly) connected component. In this case, if there exists a path of length t from a node i to itself, $(\boldsymbol{A}^t)_{ii} > 0$. The greatest common divisor (gcd) of the set $\{t : (\boldsymbol{A}^t)_{ii} > 0\}$ is called the *period* of i. If $\boldsymbol{A}$ is irreducible, then the period is the same for all indices (nodes) i. The common period is the gcd of the lengths of all the cycles in the graph. Interestingly, these topological properties of a graph have a correspondence in the spectral structure of its adjacency matrix, as shown by the Perron–Frobenius theorem (Seneta 1981).

For a nonnegative irreducible primitive matrix $\boldsymbol{A}$, the Perron–Frobenius theorem allows us to conclude that there exists an eigenvalue λ of $\boldsymbol{A}$ such that

(1) λ is real and positive, and $\lambda \geqslant |\lambda'|$ for every other eigenvalue $\lambda' \neq \alpha$;

(2) λ corresponds to a strictly positive eigenvector;

(3) λ is a simple root of the characteristic equation $(\boldsymbol{A} - \alpha \boldsymbol{I}_n) = 0$.

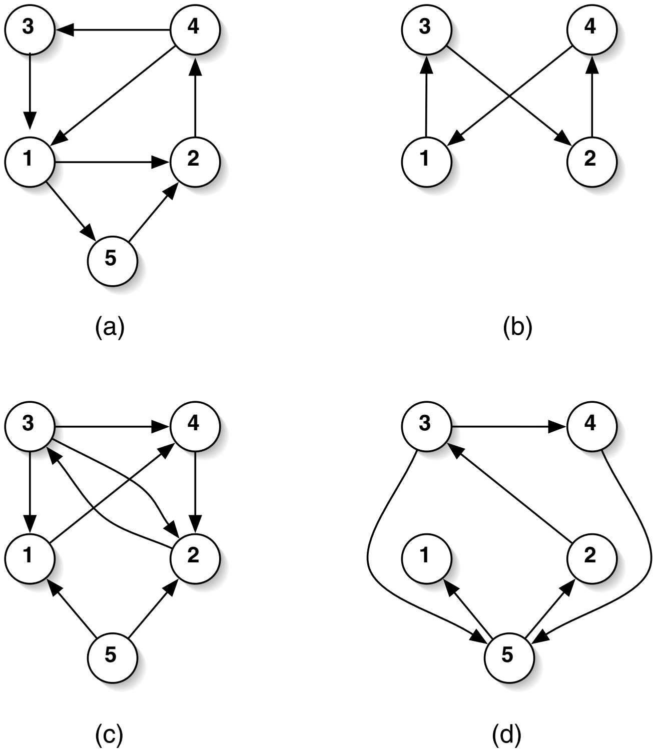

Figure 5.1 Graphs with different types of incidence matrices. (a) is primitive, (b) is irreducible (with period 4) but not primitive, (c) and (d) are reducible.

In this case, λ is called the *dominant eigenvalue* of $\boldsymbol{A}$ and the associated eigenvector is called the *dominant eigenvector*. Denoting by $(\lambda_1, \ldots, \lambda_n)$ the eigenvalues of $\boldsymbol{A}$, in the following we will assume that the dominant eigenvalue is always λ_1.

Note, however, that although there cannot be multiple roots there may be some other eigenvalue $\lambda_j \neq \lambda_1$ such that $|\lambda_j| = |\lambda_1|$. It can be shown that if there are k eigenvalues having the same magnitude as the dominant eigenvalue, then they are equally spaced in the complex circle of radius λ_1. Moreover, if $\boldsymbol{A}$ is the adjacency matrix of a graph, k is the gcd of the lengths of all the cycles in the graph. In order to get a dominant eigenvalue that is strictly greater than all other eigenvalues further conditions are necessary.

A matrix $\boldsymbol{A}$ is said to be *primitive* if there exists a positive integer t such that $\boldsymbol{A}^t > 0$ (note the strict inequality). A primitive matrix is also irreducible, but the converse is not true in general. For a primitive matrix, condition 1 of the Perron–Frobenius theorem holds with strict inequality. This means that all the remaining eigenvalues are smaller in modulus than the dominating eigenvalue. Moreover, if the adjacency matrix of a graph is primitive, then the gcd of the lengths of all cycles is unity. Figure 5.1 illustrates some examples.

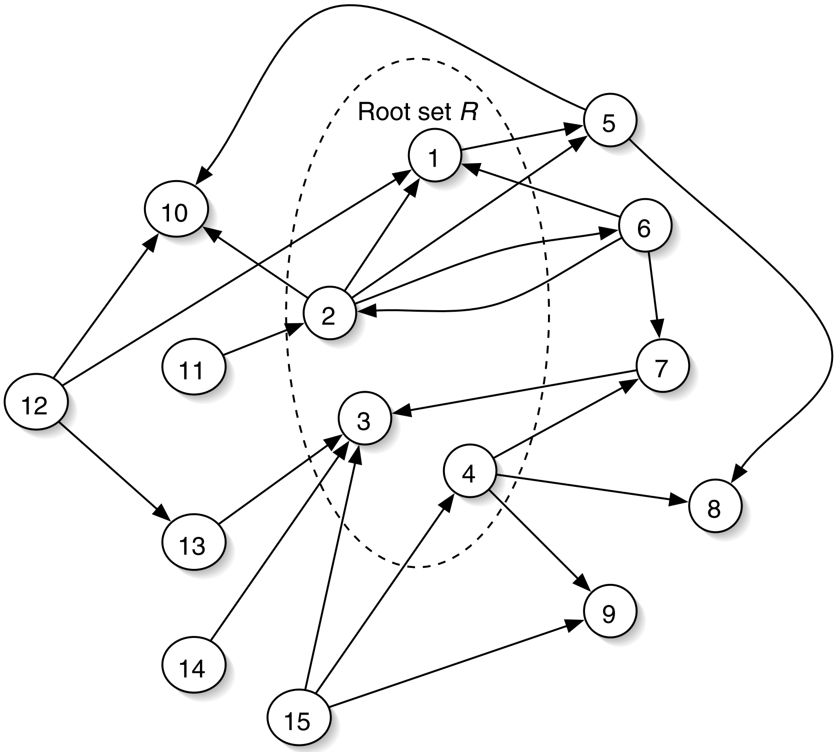

Figure 5.2 Example of a base subgraph obtainedby starting from the vertex set $\{1, 2, 3, 4\}$.

The fundamental property of primitive matrices also suggests a simple iterative algorithm for computing the dominant eigenvalue and the associated eigenvector. Let $\boldsymbol{x} \in \mathbb{R}^n$ and let $(a_1, \ldots, a_n)$ denote the coordinates of $\boldsymbol{x}$ in the basis formed by the eigenvectors $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$. If we expand the product $\boldsymbol{A}^t\boldsymbol{x}$, remembering that $\boldsymbol{A}\boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i$, we obtain

$$\boldsymbol{A}^t\boldsymbol{x} = \sum_i a_i \lambda_i^t \boldsymbol{v}_i, \tag{5.4}$$

but since $\lambda_1 > |\lambda_i|$, $i > 1$, the first term dominates the above sum as t gets large, i.e. $a_i \lambda_i \boldsymbol{v}_i \approx \boldsymbol{A}^t$ for large t. This gives us a vector proportional to the dominant eigenvector, provided that $\boldsymbol{x}$ is not orthogonal to $\boldsymbol{v}_1$. Since the Perron–Frobenius theorem tells us that $\boldsymbol{v}_1$ is strictly positive, any random positive vector will yield the correct solution, for example, $\boldsymbol{x} = \mathbf{1} = (1, 1, \ldots, 1)^{\mathrm{T}}$.

In the special case of primitive stochastic matrices, it is easy to see that $\lambda_1 = 1$ since, by the definition of a stochastic matrix, $A\mathbf{1} = \mathbf{1}$. Moreover, since all the remaining eigenvalues are strictly smaller than one in modulus, the sequence $\boldsymbol{A}^t$ converges at an exponential rate and it can be shown that

$$\lim_{t\to\infty} \boldsymbol{A}^t = \mathbf{1}^{\mathrm{T}}\boldsymbol{r}. \tag{5.5}$$

In this case, $\boldsymbol{r}$ is known as the *stationary distribution* of the Markov chain.

5.3 Hubs and Authorities: HITS

Kleinberg's algorithm, called 'Hypertext Induced Topic Selection' (HITS), simultaneously computes a pair of scoring values associated with hypertext documents (Kleinberg 1998, 1999). The semantics attached to these quantities essentially match Bray's concepts of visibility and luminosity.

HITS works on a small graph associated with a selected collection of hypertext documents, for example a focused portion of the Web that is expected to be related to a given topic of interest. In the original formulation of HITS, the subgraph of interest (also known as the *base subgraph*) is computed by selecting the neighbors of a root set R of Web pages that are known to be relevant with respect to the topic. The root set is expanded to include all the children and a fixed number of parents of nodes in R. Details of the procedure are given in the following algorithm (see Figure 5.2 for an illustration).

```
BaseSubgraph($R, d$)
1  $S \leftarrow R$
2  for each $v$ in $R$
3  do $S \leftarrow S \cup \mathrm{ch}[v]$
4     $P \leftarrow \mathrm{pa}[v]$
5     if $|P| > d$
6       then $P \leftarrow$ arbitrary subset of $P$ having size $d$
7            $S \leftarrow S \cup P$
8     return $S$
```

Note that the children of a given node (line 3) are forward links and can be obtained directly from each page v. Parents (line 4) correspond to backlinks and can be obtained from a representation of the Web graph obtained, for example, through a crawl. Several commercial search engines currently support the special query link:`url` that returns the set of documents containing `url` as a link. In the case of small scale applications, this approach can be used to obtain the set of parents in line 4. Parameter d is the maximum number of parents of a node in the root set that can be added. As we know (see Chapter 3), some pages may have a very large indegree. Thus, bounding the number of parents is crucial in practical applications. Algorithm BaseSubgraph returns a set of nodes S. In what follows, HITS is assumed to work on the subgraph of the Web induced by S.

Let $G = (V, E)$ denote the subgraph of interest, where $V = S$. For each vertex $v \in V$, let us introduce two positive real numbers $a(v)$ and $h(v)$. These quantities are called the *authority* and the *hubness* weights of v, respectively. Intuitively, a document should be very authoritative if it has received many citations. As discussed above, citations from important documents should be weighted more than citations from less-important documents. In the case of HITS, the importance of a document as a source of citations is measured by its hubness. Intuitively, a good hub is a document that allows us to reach many authoritative documents through its links. The result is that the hubness of a document depends on the authority of the cited documents, and the authority of a document depends on the hubness of the citing documents. We are apparently stuck in a loop, but let us observe that this recursive form of dependency between hubs and authority weights naturally leads to the definition of the following operations:

$$a(v) \leftarrow \sum_{w \in \mathrm{pa}[v]} h(w), \tag{5.6}$$

$$h(v) \leftarrow \sum_{w \in \mathrm{ch}[v]} a(w). \tag{5.7}$$

The two operations above can be carried out to update authority and hubness weights starting from initial values. This approach is meaningful because Kleinberg (1999) showed that iterating Equations (5.6) and (5.7), intermixed with a proper normalization step, yields a convergent algorithm. The output is a set of weights that can be therefore considered to be globally consistent. Kleinberg's algorithm is listed below. For convenience, weights are collected in two n-dimensional vectors $\boldsymbol{a}$ and $\boldsymbol{h}$.

```
HubsAuthorities($G$)
$\mathbf{1} \leftarrow [1, \ldots, 1] \in \mathbb{R}^{|V|}$
$\boldsymbol{a}_0 \leftarrow \boldsymbol{h}_0 \leftarrow \mathbf{1}$
$t \leftarrow 1$
repeat
        for each $v$ in $V$
        do $a_t(v) \leftarrow \sum_{w \in \mathrm{pa}[v]} h_{t-1}(w)$
            $h_t(v) \leftarrow \sum_{w \in \mathrm{ch}[v]} a_{t-1}(w)$
        $\boldsymbol{a}_t \leftarrow \boldsymbol{a}_t / \|\boldsymbol{a}_t\|$
        $\boldsymbol{h}_t \leftarrow \boldsymbol{h}_t / \|\boldsymbol{h}_t\|$
        $t \leftarrow t + 1$
   until $\|\boldsymbol{a}_t - \boldsymbol{a}_{t-1}\| + \|\boldsymbol{h}_t - \boldsymbol{h}_{t-1}\| < \varepsilon$
return $(\boldsymbol{a}_t, \boldsymbol{h}_t)$
```

To show that HubsAuthorities terminates, we need to prove that for each $\varepsilon > 0$ the condition controlling the outer loop will be met for t large enough. Formally, this means that the sequences $\{\boldsymbol{a}_t\}_{i \in \mathbb{N}}$ and $\{\boldsymbol{h}_t\}_{t \in \mathbb{N}}$ converge to limits $\boldsymbol{a}^\star$ and $\boldsymbol{h}^\star$, respectively. The proof of this result is based on rewriting HITS using linear algebra. In particular, if we denote by $\boldsymbol{A}$ the incidence matrix of G, it can be easily verified that the updating operations can be written compactly in vector notation as $\boldsymbol{a}_t = \boldsymbol{A}^{\mathrm{T}} \boldsymbol{h}_{t-1}$

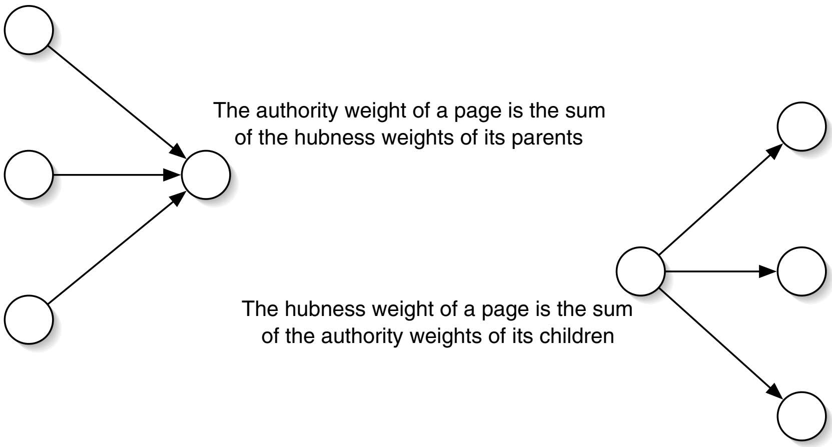
The authority weight of a page is the sum
of the hubness weights of its parents
The hubness weight of a page is the sum
of the authority weights of its children

Figure 5.3 Graphical explanation of the basic operations in HITS (see Equations (5.6) and (5.7)).

and $\boldsymbol{h}_t = \boldsymbol{A}\boldsymbol{h}_{t-1}$. As a result, after t iterations,

$$\boldsymbol{a}_t = \alpha_t(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{t-1}\boldsymbol{A}^{\mathrm{T}}\mathbf{1}, \tag{5.8}$$

$$\boldsymbol{h}_t = \beta_t(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{t}\mathbf{1}, \tag{5.9}$$

where α_t and β_t are scalar normalization factors. Thus, a sufficient condition for HubsAuthorities to terminate is that the sequences of vectors $\{\alpha_t(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{A})^{t-1}\boldsymbol{A}^{\mathrm{T}}\mathbf{1}\}$ and $\beta_t(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})^{t}\mathbf{1}$ converge for $t \to \infty$. It is possible to prove that these sequences converge under fairly unrestrictive hypotheses. For example, it can be shown that a sufficient condition is that $\boldsymbol{M}$ be a nonnegative nonsingular symmetric matrix. In this case, the dominant eigenvector $\boldsymbol{\omega}_1(\boldsymbol{M})$ (i.e. the eigenvector associated with the largest eigenvalue $\lambda_1(\boldsymbol{M})$) is nonnegative, and for every vector $\boldsymbol{x}$ such that $\boldsymbol{\omega}_1(\boldsymbol{M})^{\mathrm{T}}\boldsymbol{x} \neq 0$ we have

$$\lim_{t\to\infty} \frac{\boldsymbol{M}^t\boldsymbol{x}}{\|\boldsymbol{M}^t\boldsymbol{x}\|} \propto \boldsymbol{\omega}_1(\boldsymbol{M}). \tag{5.10}$$

Since $\mathbf{1}$ cannot be orthogonal to a nonnegative vector, the sequences $\{\boldsymbol{a}_t\}$ and $\{\boldsymbol{h}_t\}$ converge to $\boldsymbol{\omega}_1(A^{\mathrm{T}}\boldsymbol{A})$ and $\boldsymbol{\omega}_1(\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}})$, respectively.

As an example, in Figure 5.4 we show the authority and hubness weights computed by HubsAuthorities on the graph of Figure 5.2. We can note some unobvious weight assignments. For example, vertex 3 has the largest indegree in the graph but nonetheless its authority is rather small because of the low hubness weight of its parents.

Bharat and Henzinger (1998) have suggested an improved version of HITS that addresses some specific problems that are encountered in practice. For example, a mutual reinforcement effect occurs when the same host (or document) contains many identical links to the same document in another host. To solve this problem, Bharat and Henzinger (1998) modified HITS by assigning weight to these multiple edges that

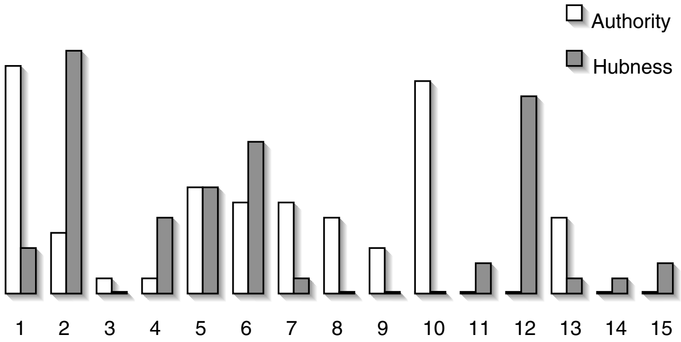

Figure 5.4 Authority and hubness weights in the example graph of Figure 5.2.

are inversely proportional to their multiplicity. The method presented by Bharat and Henzinger (1998) also addresses the problem of links that are generated automatically, for example, by converting messages posted to Usenet news groups into Web pages. Finally, they address the so-called *topic drift* problem, i.e. some nodes in the base subgraph may be irrelevant with respect to the user query and documents with highest authority or hubness weights could be about different topics. This problem can be addressed either by pruning irrelevant nodes or by regulating the influence of a node with a relevance weight.

5.4 PageRank

The theory developed in this section was introduced by Page *et al.* (1998) and resembles in many ways the recursive propagation idea we have seen in HITS. However, unlike HITS, only one kind of weight is assigned to Web documents. Intuitively, the rank of a document should be high if the sum of its parents' ranks is high. To a first approximation, this intuition might be embodied in the equation

$$r(v) = \alpha \sum_{w \in \mathrm{pa}[v]} \frac{r(w)}{|\mathrm{ch}[w]|}, \tag{5.11}$$

where $r(v)$ is the rank assigned to page v and α is a normalization constant. Note that each parent w contributes by a quantity that is proportional to its rank $r(w)$ but inversely proportional to its outdegree. This is a fundamental difference with respect to authority in HITS. The endorsement signal that flows from a given page w to each of its children decreases as the number of outgoing links (and, therefore, the potential of being a good hub) increases. Equation (5.11) can also be written in matrix notation, as

$$\boldsymbol{r} = \alpha \boldsymbol{B} \boldsymbol{r} = \boldsymbol{M} \boldsymbol{r}. \tag{5.12}$$

The matrix $\boldsymbol{B}$ is obtained from the adjacency matrix $\boldsymbol{A}$ of the graph by dividing each element by the corresponding row sum, i.e.

$$b_{uv} = \begin{cases} \dfrac{a_{uv}}{\sum_w a_{uw}}, & \text{if } \mathrm{ch}[u] \neq \emptyset, \\ a_{u,v} = 0, & \text{otherwise.} \end{cases} \tag{5.13}$$

As implied by Equation (5.12), the vector $\boldsymbol{r}$ is a right eigenvector of $\boldsymbol{B}$ with an associated eigenvalue α.

An interesting property of this solution can be seen by interpreting the computation expressed by Equation (5.12) as the description of a random walk through the Web graph. More precisely, suppose each vertex in the graph is associated with a realization of a discrete random variable S_t that models the position of a hypothetical surfer at a given time t. The rank of a page v could be then thought of as the asymptotic probability that the surfer is currently browsing that page, i.e. $P(S_t = v)$. Under this perspective, $\boldsymbol{M}$ is interpreted as the transition matrix for a first-order Markov chain:

$$\begin{aligned} r_t(v) = P(S_t = v) &= \sum_w P(S_t = v \mid S_{t-1} = w) P(S_{t-1} = w) \\ &= \sum_w m_{wv} r_{t-1}(w). \end{aligned} \tag{5.14}$$

This equation updates the probability that our random surfer will browse page v at time t, given the vector of probabilities at time $t - 1$ and the transition probabilities m_{wv}. In matrix notation this can be written as

$$\boldsymbol{r}_t = \boldsymbol{M}\boldsymbol{r}_{t-1}. \tag{5.15}$$

To satisfy probability axioms, $\boldsymbol{M}$ must be a stochastic matrix, i.e. its rows should sum to one. Since the rows of $\boldsymbol{B}$ are normalized, the probability axioms are satisfied if $\alpha = 1$. This simply means that the random surfer picks one of the outlinks in the page being visited according to the uniform distribution.

A fundamental question is whether iterating Equation (5.14) converges to some sensible solution regardless of the initial ranks $\boldsymbol{r}_0$. To answer this we need to inspect different cases in the light of the theory of nonnegative matrices developed in Section 5.2.

Four interesting cases are illustrated in Figure 5.5. The first graph has a primitive adjacency matrix. Therefore Equation (5.5) holds, and values of $r(v)$ corresponding to the steady state are indicated inside each node (below the node index). In the same figure, arcs are labeled by the amount of rank that is passed from a node to its children. As expected, Equation (5.11) holds everywhere. The second graph is more problematic, since its adjacency matrix is irreducible but not primitive. In this case, passing ranks from nodes to their children results in a cyclic updating. The random walk recursion of Equation (5.14) converges in this case to a limit cycle rather than to a steady state, and the periodicity of the limit cycle is the period of the matrix, or, as

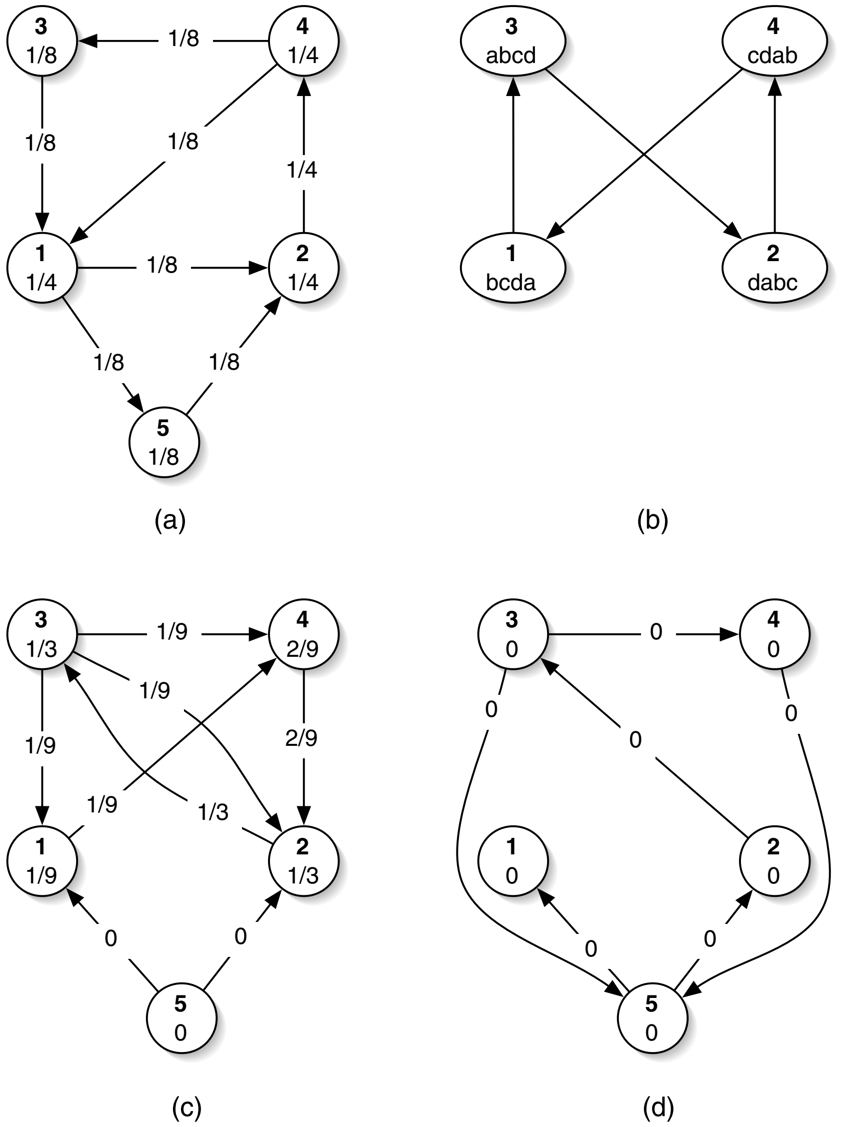

Figure 5.5 Rank propagation on graphs with different types of incidence matrices. Equation (5.14) converges to a nontrivial steady state in case (a) and (c), to a limit cycle in case (b), and to zero in case (d).

we know, the gcd of the lengths of the cycles (4 in the example). This is indicated in Figure 5.5b by four values a, b, c, d of rank that cyclically bounce along the nodes. The third graph of Figure 5.5 has a reducible adjacency matrix. In this case $\boldsymbol{M}^t$ converges to a matrix whose last column is all zero, reflecting the fact that the node should have zero rank as it has no parents. Finally, the fourth graph has also a reducible adjacency matrix but this time the maximum eigenvalue is less than one, so $\boldsymbol{M}^t$ converges to the zero matrix. This is due to the existence of a node (1) with no children that effectively acts as a rank sink.

The situation in Figure 5.5d is of course undesirable but is very common in the actual Web. Many pages have no outlinks at all. Furthermore, pages that remain on the crawling frontier and are never fetched will likely produce dangling edges in the graph that is obtained from crawling. To solve this difficulty, observe that the connectivity of node 1 should be defined as illegal, since it violates the basic hypothesis underlying the random walk model: the sum of the probabilities of the available actions should be one in each node, but once in the sink nodes our random surfer would be left with no choices. A sensible correction consists of giving the random surfer a 'method of escape' by adding allowable actions. One possibility is to assume that the surfer, who cannot possibly follow any link, will restart browsing by picking a new Web page at random. This is the same as adding a link from each sink to each other vertex, i.e. introducing an escape matrix $\boldsymbol{E}$ defined as $e_{vw} = 0$ if $|\mathrm{ch}[v]| > 0$ and $e_{vw} = 1/n$ otherwise, for each w. Then the transition matrix becomes

$$\boldsymbol{M} = (\boldsymbol{B} + \boldsymbol{E}).$$

$\boldsymbol{M}$ is now a stochastic matrix and the Markov chain model for a Web surfer is sound. In general, however, there is no guarantee that $\boldsymbol{M}$ is also primitive (if there are cycles with zero outdegree as in Figure 5.1b, these bring irreducible but periodic components). This difficulty will be addressed shortly and for now let us assume that $\boldsymbol{M}$ is primitive.

The following iterative algorithm was suggested in the original paper on PageRank (Page *et al.* 1998). It takes as input a nonnegative square matrix M, its size n, and a tolerance parameter ϵ.

```
PageRank($\boldsymbol{M}, n, \epsilon$)
$\mathbf{1} \leftarrow [1, \dots, 1] \in \mathbb{R}^n$
$z \leftarrow \frac{1}{n}\mathbf{1}$
$\boldsymbol{x}_0 \leftarrow z$
$t \leftarrow 0$
repeat
        $t \leftarrow t + 1$
        $\boldsymbol{x}_t \leftarrow \boldsymbol{M}^{\mathrm{T}}\boldsymbol{x}_{t-1}$
        $d_t \leftarrow \|\boldsymbol{x}_{t-1}\|_1 - \|\boldsymbol{x}_t\|_1$
        $\boldsymbol{x}_t \leftarrow \boldsymbol{x}_1 + d_t z$
        $\delta \leftarrow \|\boldsymbol{x}_{t-1} - \boldsymbol{x}_t\|_1$
   until $\delta < \epsilon$
return $\boldsymbol{x}_t$
```

The quantity d_t is the total rank being lost in sinks. Adding $d_t z$ to $\boldsymbol{M}^{\mathrm{T}}\boldsymbol{x}_{t-1}$ is basically a normalization step. As it turns out, if $\boldsymbol{M}$ is a stochastic primitive matrix, then $d_t = 0$ in each iteration (no normalization is necessary) and PageRank converges to the stationary distribution of $\boldsymbol{M}$. Otherwise, the above algorithm implicitly 'repairs' the matrix $\boldsymbol{M}$ into a stochastic matrix and converges to the corresponding stationary distribution (see Exercise 5.4).

Now we address the problem of irreducible but periodic components. These also act as rank sinks because they never pass rank to other parts of the graph. Moreover, periodicity may hurt convergence and the algorithm PAGERANK above is not anymore guaranteed to terminate. The solution suggested in Page *et al.* (1998) consists of forcing some source or rank by introducing a 'static' stochastic process that models the 'distribution of Web pages that a random surfer periodically jumps to.' This distribution can be any nonnegative vector $\boldsymbol{e}$ such that $\|\boldsymbol{e}\|_1 = 1$. The probability distribution that results from combining the Markovian random walk distribution $\boldsymbol{x}$ and the static rank source distribution is a mixture model with parameter ε:

$$\boldsymbol{r} = \varepsilon \boldsymbol{e} + (1 - \varepsilon)\boldsymbol{x}.$$

The simplest choice for $\boldsymbol{e}$ is a uniform distribution, i.e. $\boldsymbol{e} = (1/n)\mathbf{1}$. Intuitively, this approach can be motivated by the metaphor that browsing consists of following existing links with some probability $1-\varepsilon$ or selecting a nonlinked page with probability ε. When the latter choice is made, each page in the entire Web is sampled according to the probability distribution $\boldsymbol{e}$. In the case of the uniform distribution, Equation (5.15) will be rewritten as

$$\boldsymbol{r}_t = [\varepsilon \boldsymbol{H} + (1 - \varepsilon)\boldsymbol{M}]^{\mathrm{T}} \boldsymbol{r}_{t-1}, \tag{5.16}$$

where $\boldsymbol{H}$ is a square matrix with $h_{uv} = 1/n$ for each u, v. In this way we have obtained an ergodic Markov chain whose underlying transition graph is fully connected. The associated transition matrix $\varepsilon \boldsymbol{H} + (1-\varepsilon)\boldsymbol{M}$ is primitive and therefore the sequence $\boldsymbol{r}_t$ converges to the dominant eigenvector. The stationary distribution $\boldsymbol{r}$ associated with the Markov chain described by Equation (5.16) is known as *PageRank*. In practice, ε is typically chosen to be between 0.1 and 0.2 (Brin and Page 1998).

5.5 Stability

An important question is whether the link analysis algorithms based on eigenvectors (such as HITS and PageRank) are stable in the sense that results do not change significantly as a function of modest variations in the structure of the Web graph. More precisely, suppose the connectivity of a portion of the graph is changed arbitrarily, i.e. let $G = (V, E)$ be the graph of interest and let us replace it by a new graph $\tilde{G} = (V, \tilde{E})$, where some edges have been added or deleted. How will this affect the results of algorithms such as HITS and PageRank?

Ng *et al.* (2001) proved two interesting results about the stability of algorithms based on the computation of dominant eigenvectors.

5.5.1 Stability of HITS

First, Ng *et al.* (2001) derived a bound on the number of hyperlinks that can added or deleted from one page without significantly affecting the authority (or hubness) weights computed by HITS. The bound essentially depends on the *eigengap*, namely the difference $\delta \doteq \lambda_1 - \lambda_2$ between the two largest eigenvalues of $\boldsymbol{M} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ (since this matrix is symmetric the eigengap is a real number).

The result can be formally stated as follows. For every $\alpha > 0$, suppose G is perturbed by adding or deleting at most k hyperlinks from one page,

$$k \leqslant \left(\sqrt{d + \frac{\alpha\delta}{4 + \sqrt{2}\alpha}} - \sqrt{d}\right)^2, \tag{5.17}$$

where d is the maximum outdegree of G. The principal eigenvector associated with the perturbed graph $\tilde{G}$ then satisfies

$$\|\boldsymbol{a} - \tilde{\boldsymbol{a}}\|_2 \leqslant \alpha. \tag{5.18}$$

Moreover, Ng *et al.* (2001) show that it is possible to perturb a symmetric matrix, by a quantity that grows as δ, that produces a constant perturbation of the dominant eigenvector. Thus, matrices with small eigengap can have low robustness with respect to perturbation.

In practice, it is not difficult to construct graphs where even adding or deleting even a single edge results in large variations in the authority and hubness weights. For example, consider two isolated communities, one possessing a hub h but no emerging authority (so the authority weight is dispersed among all the nodes), and the other possessing a well recognized authority a but no important hubs. Adding an edge from h to a would result in a large change in weight assignments, since the authority weight of a would be increased at the expense of some amount of authority weights contributed by all the nodes in the first community.

5.5.2 Stability of PageRank

In this case suppose $\boldsymbol{r}$ is the stationary distribution associated with matrix

$$\varepsilon\boldsymbol{H} + (1 - \varepsilon)\boldsymbol{M}$$

(see Equation (5.16)). If the adjacency matrix $\boldsymbol{A}$ is perturbed to a new matrix $\tilde{\boldsymbol{A}}$, then Ng *et al.* (2001) show that

$$\|\tilde{\boldsymbol{r}} - \boldsymbol{r}\| \geqslant \frac{2\sum_{j\in\tilde{V}} r(j)}{\varepsilon}, \tag{5.19}$$

where $\tilde{V}$ denotes the set of vertices touched by the perturbation. This demonstrates two interesting facts: first, the parameter ε of the mixture model in Equation (5.16) has a stabilization role; second, if the set of pages affected by the perturbation have a

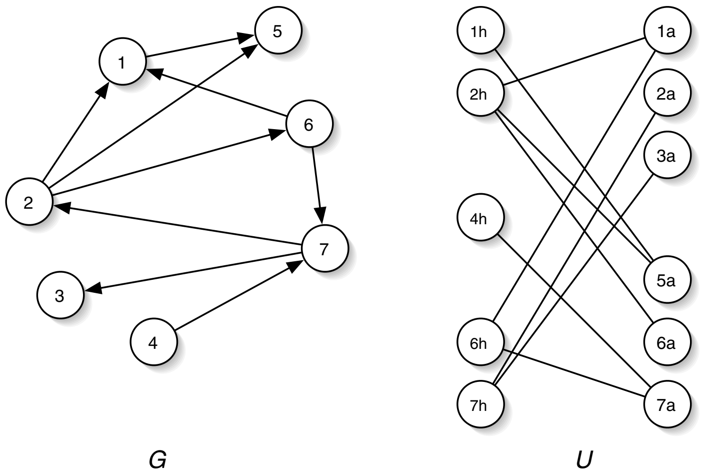

Figure 5.6 Forming a bipartite graph in SALSA.

small rank, the overall change will also be small. Bianchini *et al.* (2001) later proved the tighter bound

$$\|\tilde{\boldsymbol{r}} - \boldsymbol{r}\| \geqslant \frac{1-\varepsilon}{\varepsilon} 2 \sum_{j \in \tilde{V}} \delta(j) r(j), \tag{5.20}$$

where $\delta(j) \geqslant 2$ depends on the edges incident on j affected by the perturbation.

5.6 Probabilistic Link Analysis

The probabilistic interpretation of PageRank based on random walks can be extended to link analysis algorithms that, like HITS, distinguish the importance of a node as an authority or as a hub.

5.6.1 SALSA

Lempel and Moran (2001) have proposed a probabilistic extension of HITS called the ‘Stochastic Approach for Link Structure Analysis’ (SALSA). Similar extensions have been proposed independently by Rafiei and Mendelzon (2000) and Ng *et al.* (2001). In all of these proposals, the random walk is carried out by following hyperlinks both in the forward and in the backward direction.

SALSA starts from a graph $G = (V, E)$ of topically related pages (like the base subgraph of HITS) and constructs a bipartite undirected graph $U = (\hat{V}, \hat{E})$ as (see Figure 5.6)

$$\hat{V} = V_{\mathrm{h}} \cup V_{\mathrm{a}},$$

where

$$V_{\mathrm{h}} \doteq \{v_{\mathrm{h}} : v \in V,\ \mathrm{ch}[v] \neq \emptyset\},$$
$$V_{\mathrm{a}} \doteq \{v_{\mathrm{a}} : v \in V,\ \mathrm{pa}[v] \neq \emptyset\},$$
$$\hat{E} \doteq \{(u_{\mathrm{h}}, v_{\mathrm{a}}) : (u, v) \in E\}.$$

The sets V_{h} and V_{a} are called the *hub side* and the *authority side* of U, respectively.

Two separate random walks are then introduced. In the 'hub' walk, each step consists of

(1) following a Web link from a page u_{h} to a page w_{a}, and

(2) immediately afterward following a backlink going back from w_{a} to v_{h}, where we have assumed that $(u, w) \in E$ and $(v, w) \in E$.

For example, jumping from 1_{h} to 5_{a} and then back from 5_{a} to 2_{h} in Figure 5.6. In the 'authority' walk, a step consists of following a backlink first and a forward link next. In both cases, a step translates into following a path of length exactly two in U. Note that, by construction, each walk starts on one side of U, either the hub side or the authority side, and will remain confined to the same side. The Markov chains associated with the two random walks have transition matrices $\tilde{\boldsymbol{H}}$ and $\tilde{\boldsymbol{T}}$, respectively, defined as follows:

$$\tilde{h}_{uv} = \sum_{\substack{w:(u,w)\in E,\\ (v,w)\in E}} \frac{1}{\deg(u_{\mathrm{h}})} \frac{1}{\deg(w_{\mathrm{a}})},$$

$$\tilde{t}_{uv} = \sum_{\substack{w:(w,u)\in E,\\ (w,v)\in E}} \frac{1}{\deg(v_{\mathrm{a}})} \frac{1}{\deg(w_{\mathrm{h}})}.$$

The hub and authority weights are then obtained as principal eigenvectors of the matrices $\tilde{\boldsymbol{H}}$ and $\tilde{\boldsymbol{T}}$. Note that these two matrices could also be defined in an alternative way. Let $\boldsymbol{A}$ be the adjacency matrix of G, $\boldsymbol{A}_{\mathrm{r}}$ the row-normalized adjacency matrix (as in Equation (5.13)) and let $\boldsymbol{A}_{\mathrm{c}}$ the column-normalized adjacency matrix of G (i.e. dividing each nonzero entry by its column sum). Then $\tilde{\boldsymbol{H}}$ consists of the nonzero rows and columns of $\boldsymbol{A}_{\mathrm{r}} \cdot \boldsymbol{A}_{\mathrm{c}}^{\mathrm{T}}$, while $\tilde{\boldsymbol{T}}$ consists of the nonzero rows and columns of $\boldsymbol{A}_{\mathrm{c}}^{\mathrm{T}} \cdot \boldsymbol{A}_{\mathrm{r}}$.

Note that $\tilde{h}_{uv} > 0$ implies that there exists at least one page w that has links to both u and v. This is known as co-citation in bibliometrics (Kessler 1963) (see Figure 5.9 and Exercise 5.2). Similarly, $\tilde{t}_{uv} > 0$ implies there exists at least one page that is linked to by both u and v, a bibliographic coupling (Small 1973).

Lempel and Moran (2001) showed theoretically that SALSA weights are more robust that HITS weights in the presence of the Tightly Knit Community (TKC) Effect. This effect occurs when a small collection of pages (related to a given topic) is connected so that *every* hub links to *every* authority and includes as a special case the mutual reinforcement effect identified by Bharat and Henzinger (1998) (see

Section 5.3). It can be shown that the pages in a community connected in this way can be ranked highly by HITS, higher than pages in a much larger collection where only *some* hubs link to *some* authorities. Clearly the TKC effect could be deliberately created by spammers interested in pushing the rank of certain websites. Lempel and Moran (2001) constructed examples of community pairs C_s connected in a TKC fashion, and C_l sparsely connected, and proved that authorities of C_s are ranked above the authorities of C_l by HITS but not by SALSA.

In a similar vein, Rafiei and Mendelzon (2000) and Ng *et al.* (2001) have proposed variants of the HITS algorithm based on a random walk model with reset, similar to the one used by PageRank. More precisely, a random surfer starts at time $t = 0$ at a random page and subsequently follows links from the current page with probability $1 - \varepsilon$, or (s)he jumps to a new random page with probability ε. Unlike PageRank, in this model the surfer will follow a forward link on odd steps but a backward link on even steps. For large t, two stationary distributions result from this random walk, one for odd values of t, that corresponds to an authority distribution, and one for even values of t that correspond to a hubness distribution. In vector notation the two distributions are proportional to

$$\boldsymbol{a}_{2t+1} = \varepsilon\mathbf{1} + (1 - \varepsilon)\boldsymbol{A}_{\mathrm{r}}\boldsymbol{h}_{2t}, \tag{5.21}$$

$$\boldsymbol{h}_{2t} = \varepsilon\mathbf{1} + (1 - \varepsilon)A_{\mathrm{c}}^{\mathrm{T}}\boldsymbol{a}_{2t-1}. \tag{5.22}$$

The stability properties of these ranking distributions are similar to those of PageRank (Ng *et al.* 2001).

Some further improvements of HITS and SALSA, as well as theoretical analyses on the properties of these algorithms can be found in Borodin *et al.* (2001).

5.6.2 PHITS

Cohn and Chang (2000) point out a different problem with HITS. Since only the principal eigenvector is extracted, the authority along the remaining eigenvectors is completely neglected, despite the fact that it could be significant. An obvious approach to address this limitation consists of taking into account several eigenvectors of the co-citation matrix, in the same spirit as PCA is used to extract several factors that are responsible for variations in multivariate data. As we have discussed in Section 4.5.2, however, the statistical assumptions underlying PCA are not sound for multinomial data such as term–document occurrences or bibliographical citations. PHITS can be viewed as probabilistic LSA (see Section 4.5.2) applied to co-citation and bibliographic coupling matrices. In this case citations replace terms. As in PLSA, a document d is generated according to a probability distribution $P(d)$ and a 'latent' variable z is then attached to d with probability $P(z \mid d)$. Here z could represent research areas (in the case of bibliographic data) or a (topical) community in the case of Web documents. Citations (links) are then chosen according to a probability distribution $P(d' \mid z)$.

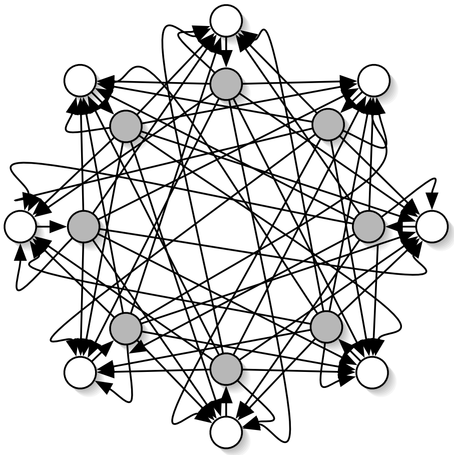

Figure 5.7 A link farm. Shaded nodes are all copies of the same page.

5.7 Limitations of Link Analysis

Search engines can be 'spammed' by websites faking high relevance with respect to some topics for the sole purpose of attracting visitors. Marchiori (1997) called this phenomenon Search Engine Persuasion (SEP). First-generation engines that based their ranking on classical information retrieval measures were clearly very sensitive to SEP. Common early techniques used to fool these engines included the use of inappropriate site titles or descriptions, the inclusion of META keywords in the HTML code, or even the use of extra text invisible to human surfers. Link analysis was immediately recognized as a solid defense against SEP. In their seminal paper on PageRank, Page *et al.* (1998) stated that

> . . . for a page to get a high PageRank, it must convince an important page, or a lot of non-important pages to link to it. At worst, you can have manipulation in the form of buying advertisements (links) on important sites. But this seems well under control since it costs money. This immunity to manipulation is an extremely important property.

In the intervening time period, ranking pages using link analysis has become the standard approach followed by all the 'second generation' search engines. Consequently, website owners have realized the enormous importance of maximizing PageRank or similar link-based scoring functions in order to increase visibility. For some websites, having high search engine ranking (with respect to some keywords) is so important that it justifies considerable financial investments. Buying links as a form of advertisement has become reality and perhaps in some cases it is a sensible alternative to paying search engine companies directly for advertised links.

A *link farm* is a densely connected Web subgraph artificially built for the purpose of accumulating PageRank (or similar measures of popularity). Link farms can be built in many ways but link exchange is a common approach. A website owner who

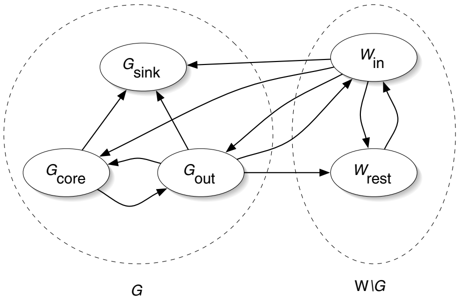

Figure 5.8 Canonical decomposition of the Web relative to a subgraph G.

decides to join the farm agrees to store a copy of a 'hub' page on her server and to link it from the root of her site. In return, the main URL of her site is added to the hub page, which is in turn redistributed to the sites participating in the link exchange. The result is a densely connected subgraph like the one shown in Figure 5.7.

It may appear that, since structures of this kind are highly regular, they should be relatively easy to detect (see Exercise 5.8) and thus link farming should not be a serious concern for search engines. However, it is possible to build farms that are more tightly entangled in the Web and are therefore more difficult to detect by simple topological analyses. This problem has been recently pointed out by Bianchini *et al.* (2001), who have shown that every community, defined as an arbitrary subgraph G of the Web, must satisfy a special form of 'energy balance'. The overall PageRank assigned to pages in G grows with the 'energy' that flows in from pages linking to the community and decreases with the energy dispersed in sinks and passed to pages outside the community. With reference to Figure 5.8, let G_{out} denote the subgraph of G induced by pages that contain hyperlinks pointing outside to G and let G_{sink} denote the sink subgraph of G. Also, let W_{in} be the subgraph induced by the pages outside G that link to pages in G. Then the equation

$$\|\boldsymbol{r}_G\| = \alpha|G| + E_G^{\text{in}} - E_G^{\text{sink}} - E_G^{\text{out}}. \tag{5.23}$$

holds, where $\alpha|G|$ is the 'default' energy that is assigned to the community,

$$E_G^{\text{in}} = \frac{1-\varepsilon}{\varepsilon} \sum_{w \in W_{\text{in}}} f_G(w)\, r(w)$$

is the energy flowing in from outside, where $f_G(w)$ is the fraction of links in w that point to pages in G,

$$E_G^{\text{out}} = \frac{1-\varepsilon}{\varepsilon} \sum_{w \in G_{\text{out}}} (1 - f_G(w))\, r(w)$$

is the energy flowing out to pages outside the community, and

$$E_G^{\text{sink}} = \frac{1-\varepsilon}{\varepsilon} \sum_{w \in G_{\text{sink}}} r(w).$$

Suppose a set of ‘sponsoring’ pages S is generated artificially to boost the rank of a target website. The above energy balance analysis shows that the PageRank of the target can be increased linearly with $|S|$ regardless of the topology in S, making it difficult to detect the origin of spam using methods that are only based on graph topology. The fact that spamming activities of this kind are indeed possible is borne out, for example, by some recent anecdotes such as the popularity battle of the Church of Scientology against its main opponent Xenu.net.[2]

Websites that allow their users to post HTML code, for example, Weblogs (Walker 2002), potentially offer a cheap way for constructing artificial rank-boosting communities. Specialized algorithms that exploit more information than just topology are likely to be needed in the near future in order to prevent these spamming activities. This may also help prevent the future topology of the Web (and associated notions of popularity) from becoming significantly controlled by economic interests.

One example in this direction is the paper by Davison (2000a) that describes a machine-learning method for the automatic discrimination of links that are unrelated to the intrinsic merit of the pages. In this case, hyperlinks are described by several binary features that include, for example, tests about the structure of the URL, identity of source and target host or domain, or the total number of outlinks found on the source page.

Finally, we note that assessing the *quality* of pages returned by a search engine is difficult because quality is ultimately defined by human judgement. Amento *et al.* (2000) have reported an empirical study in which 16 human experts in five popular topics related to TV entertainment and music were asked to rank the quality of a set of Web pages. The experiment was aimed at testing whether expert judgements were correlated with scores based on link analysis. The study revealed that ranking documents according to the authority score (as computed by HITS) or according to PageRank yields high precision for the top 5 or 10 ranked documents. However, it was also found that alternative metrics such as page indegree or total number of pages in the website perform equally well.

Exercises

Exercise 5.1. Draw a graph of reasonable size, connecting vertices at random, and compute the principal eigenvectors of the matrices $\boldsymbol{A}^{\mathrm{T}}A$ and $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ to get authority and hubness weights. A very rapid way of doing this is by using linear algebra software such as Octave. Now select a vertex having nonzero indegree but small authority and try to modify the graph to increase its authority without increasing its indegree nor the indegree of its parents.

Exercise 5.2. The two matrices involved in Equations (5.8) and (5.9) were introduced several years before in the field of bibliometrics. In particular, $\boldsymbol{C} = \boldsymbol{A}^{\mathrm{T}}\boldsymbol{A}$ is known as the *co-citation matrix* (Kessler 1963) and $\boldsymbol{B} = \boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ is known as the *bibliographic*

[2] See `http://www.operatingthetan.com/google/` for details.

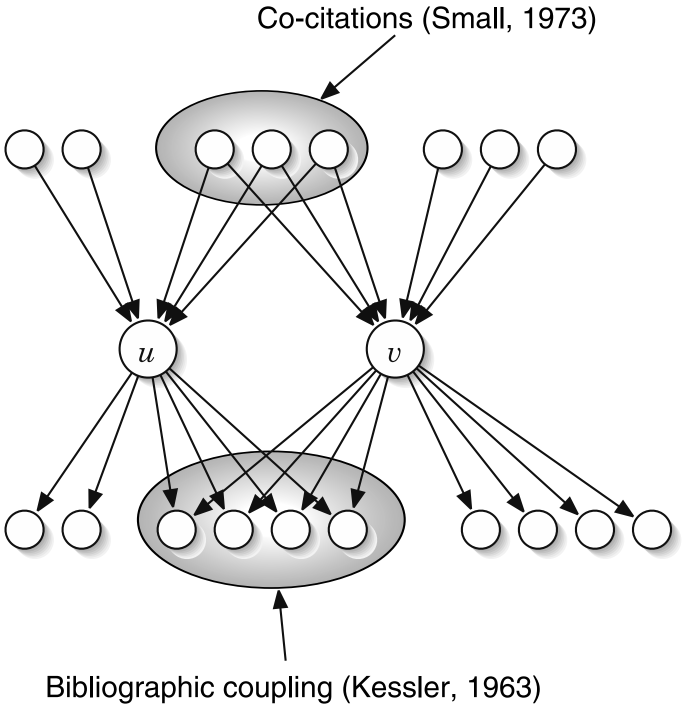
Co-citations (Small, 1973)
u
v
Bibliographic coupling (Kessler, 1963)

Figure 5.9 Co-citations and bibliographic coupling.

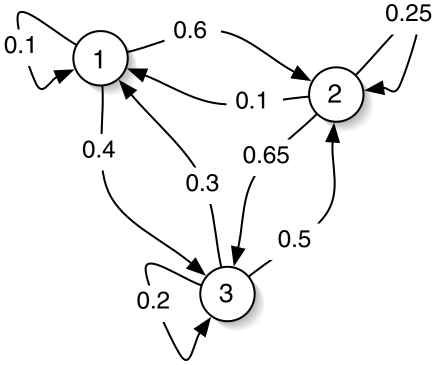

Figure 5.10 Markov chain for Exercise 5.3.

coupling matrix (Small 1973). Show that c_{uv} is the number of documents that cite both documents u and v, while b_{uv} is the number of pages that are cited by both u and v (see Figure 5.9).

Exercise 5.3. Consider the Markov chain in Figure 5.10 (where arcs are labeled by transition probabilities). Is it ergodic? What is the steady-state distribution?

Exercise 5.4. Let $\boldsymbol{N}$ be the $n \times n$ row normalized adjacency matrix of a graph as in Section 5.4. Suppose $\boldsymbol{N}$ is not stochastic and let $\boldsymbol{R}$ be any matrix such that $\boldsymbol{N} + \boldsymbol{R}$ is a stochastic matrix. Show that $\textsc{PageRank}(\boldsymbol{N}, n, \epsilon)$ and $\textsc{PageRank}(\boldsymbol{N} + \boldsymbol{R}, n, \epsilon)$ converge to the same solution.

Exercise 5.5. In Equation (5.21), the stationary distributions of authority and hubness for randomized HITS are defined within a proportionality constant. Determine what this constant is.

Exercise 5.6. Use the stability results in Section 5.5.2 to estimate how often you should crawl the Web (and recompute PageRank) in order to guarantee that

$$\|\tilde{\boldsymbol{r}} - \boldsymbol{r}\| < \epsilon,$$

ϵ being an assigned tolerance. Assume for simplicity that a constant number of pages are changed in a given unit of time.

Exercise 5.7. Implement the PageRank computation and simulate the results on a relatively large artificial graph (build the graph using ideas from Chapter 3). Then introduce link farms in your graph and study the effect they have on the PageRank vector as a function of the number and the size of the farms.

Exercise 5.8. Propose an efficient algorithm to detect link farms structured as in Figure 5.7 in a large graph.